On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection

Xiufeng Song[1], **Xiao Guo**[2], **Jiache Zhang**[1], **Qirui Li**[1],
Lei Bai[3], **Xiaoming Liu**[2], **Guangtao Zhai**[1], **Xiaohong Liu**[*1]
[1]Shanghai Jiao Tong University [2]Michigan State University [3]Shanghai AI Laboratory
{akikaze, zjc_he, iapple1, zhaiguangtao, xiaohongliu}@sjtu.edu.cn
{guoxia11, liuxm}@cse.msu.edu baisanshi@gmail.com
[*] Corresponding Author

Abstract

Large numbers of synthesized videos from diffusion models pose threats to information security and authenticity, leading to an increasing demand for generated content detection. However, existing video-level detection algorithms primarily focus on detecting facial forgeries and often fail to identify diffusion-generated content with a diverse range of semantics. To advance the field of video forensics, we propose an innovative algorithm named Multi-Modal Detection(MM-Det) for detecting diffusion-generated videos. MM-Det utilizes the profound perceptual and comprehensive abilities of Large Multi-modal Models (LMMs) by generating a Multi-Modal Forgery Representation (MMFR) from LMM's multi-modal space, enhancing its ability to detect unseen forgery content. Besides, MM-Det leverages an In-and-Across Frame Attention (IAFA) mechanism for feature augmentation in the spatio-temporal domain. A dynamic fusion strategy helps refine forgery representations for the fusion. Moreover, we construct a comprehensive diffusion video dataset, called Diffusion Video Forensics (DVF), across a wide range of forgery videos. MM-Det achieves state-of-the-art performance in DVF, demonstrating the effectiveness of our algorithm. Both source code and DVF are available at link.

1 Introduction

Recent years have witnessed significant advancements in diffusion generative methods, which have led to the creation of extraordinarily visually compelling content in video generation [5, 4, 61]. Although the latest generated videos impress society with their versatility and stability, synthetic media also poses a risk of malicious attacks, such as counterfeit faces created by deepfakes [49] and falsifications in business, raising public concerns about information security and privacy. In response to such issues, researchers have made significant progress in forgery detection, addressing problems on image editing manipulation [60, 30, 17] and CNN-synthesized images [53, 40, 55, 36, 18]. To enhance the trustworthiness and reliability of current detectors in the face of evolving generative video methods, we aim to develop a generalizable detection method for diffusion-based generative videos.

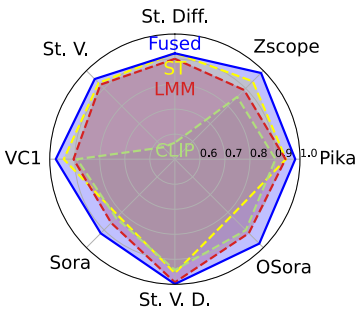


Figure 1: Multi-Modal Detection (MM-Det) leverages features from spatiotemporal (ST) information (), a CLIP encoder [39] (), and an LMM (). The Fused feature () achieves state-of-the-art performance in our Diffusion Video Forensics (DVF) dataset.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

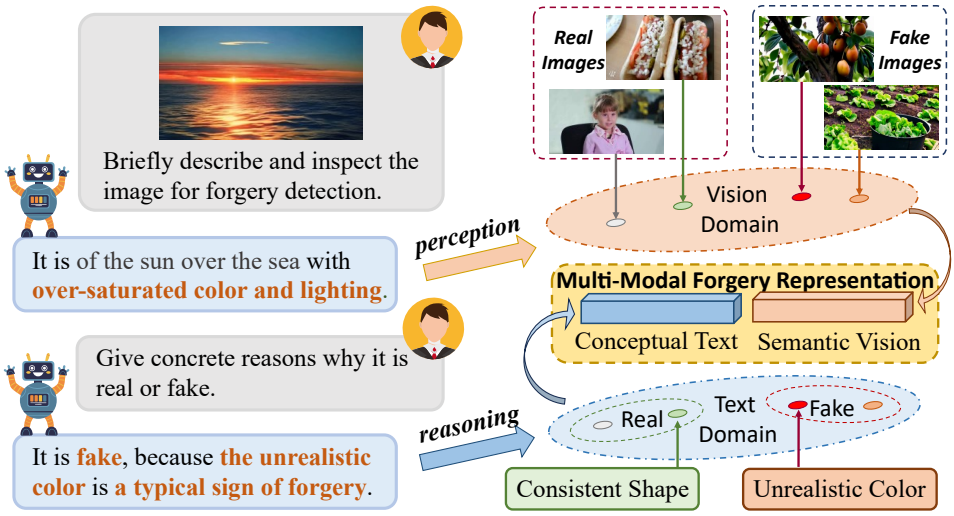


Figure 2: LMMs detect visual artifacts and anomalies, offering detailed textual reasoning that explains whether the image is generated using Artificial Intelligence (AI) techniques. The powerful representation in the visual domain enables LMMs to understand complex contexts within frames. Furthermore, their advanced language reasoning capability implicitly reveals image authenticity and provenance. For instance, the term like ‘‘`consistent shape`’’ refers to common features in authentic content, while ‘‘`unrealistic color`’’ signifies typical artifacts in forged content. This linguistic proficiency stems from the superior perception and comprehension abilities of LMMs, contributing to a generalizable multimodal feature space. By leveraging the visual understanding and textual reasoning abilities of LMMs, we construct a Multi-Modal Forgery Representation (MMFR).

Previously, the video forensics community emphasized more on developing facial forgery detection algorithms [57, 73, 10, 62, 70], which may struggle to address recent fraudulent videos (*e.g.*, sora, pika, etc.). Compared to facial forgery, diffusion-based generated content contains more diverse semantics, making it more challenging to distinguish diffusion forgery contents from real ones. Towards these challenges, a new thread of research on CNN-generated image detection has emerged [53, 40, 55, 36, 18]. These works aim to learn common generation traces in image-level content, but do not design specific mechanisms to capture temporal inconsistencies in videos.

Therefore, previous defensive efforts might not be able to provide a video-level detection algorithm for newly emerged generated videos with diverse manipulation artifacts and visual contexts. Meanwhile, Large Multi-modal Models (LMMs) show unparalleled problem-solving ability [2, 24, 23, 75, 69, 28], thanks to its powerful multi-modal representations. However, such representations are barely studied in the video forensics task.

Motivated by the limitation of previous work and the unprecedented understanding ability of LMMs, we propose a video-level detection algorithm, named Multi-Modal Detection (MM-Det), to capture forgery traces based on an LMM-based multi-modal representation. MM-Det takes advantage of the perception and reasoning capabilities of LMMs to learn a generalizable forgery feature, as depicted in Fig. 2. To the best of our knowledge, we are the first to use LMMs for video forensic work.

Aside from multi-modal representations, two common sources of generative errors that can be leveraged for video discrimination are spatial artifacts and temporal inconsistencies. Our approach aims to effectively identify these two types of errors as an auxiliary feature in forgery detection. Inspired by the previous work [55, 33, 41] that shows the effectiveness of reconstruction for detecting diffusion images, we extend this idea into the video domain, amplifying diffusion artifacts both in spatial and temporal information. To capture such artifacts efficiently, we leverage a Vector Quantised-Variational AutoEncoder (VQ-VAE) [51] for a fast reconstruction process, as detailed in Fig. 3. Moreover, we design a novel **I**n-and-**A**cross **F**rame **A**ttention (IAFA) into a Transformer-based network, which balances frame-level forgery traces with information flow across frames, thus aggregating local and global features.

Although diffusion methods demonstrate strong capabilities in video generation, the lack of public datasets on diffusion videos hinders research efforts in the video forensic community. In light of this,

we have established a comprehensive dataset for diffusion-generated videos, named Diffusion Video Forensics (DVF). DVF includes generated content from a variety of diffusion models, featuring rich semantics and high quality, serving as a general benchmark for open-world video forensics tasks. The main contributions of this paper are as follows:

◇ We propose a detection method called MM-Det that leverages a **M**ulti-**M**odal Forgery Representation from LMMs to effectively detect diffusion-generated videos with strong generalization capability.

◇ A powerful and innovative In-and-Across Frame Attention (IAFA) mechanism is introduced to aggregate global and local patterns within forged videos, enhancing the detection of spatial artifacts and temporal inconsistencies.

◇ We introduce a large-scale dataset, named the Diffusion Video Forensics (DVF) dataset, comprising high-quality forged videos generated using 8 diffusion-based methods. The DVF dataset contains diverse forgery types across videos of varying resolutions and durations, effectively serving as a benchmark for forgery detection in real-world scenarios.

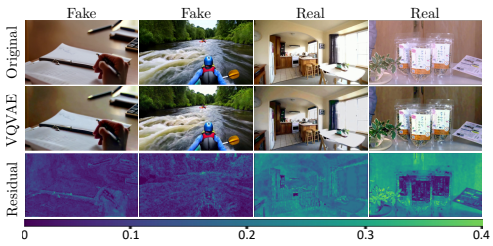


Figure 3: The residual difference between VQ-VAE [51] reconstructed images and real ones. Given an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ of a VQ-VAE and taking the input video $\mathbf{v}$, the reconstructed video $\mathbf{v}'$ is obtained as $\mathbf{v}' = \mathcal{D}(\mathcal{E}(\mathbf{v}))$. The VQ-VAE reconstruction of real images exhibits obvious edges and visible traces, whereas diffusion-generated ones are reconstructed more effectively, offering residual difference images with fewer visible traces.

◇ Our MM-Det achieves state-of-the-art detection performance on the DVF dataset. Also, a detailed analysis is provided to showcase the effectiveness of multi-modal representations in detecting forgeries, paving the way for compelling opportunities for using LMMs in future multi-media forensic research.

2 Related Works

Frame-level Detector Early work [53, 22, 13, 47, 65, 16] observed that forgery traces exist in images generated by AI techniques, and such traces are commonly used as evidence to distinguish diffusion-generated content [40, 6, 7] and attribute if two images are generated by the same method [66, 37]. However, identifying unseen and diverse frequency-based clues in real-world scenarios is challenging. For that, existing frame-level forgery detectors concentrate on improving the generalization ability. For example, some works [36, 8, 17, 26, 29] introduced features from pre-trained CLIP [39] encoders for the forensic task to help reduce the overfitting issue on specific forgery types and increase the robustness towards detection [36, 8] and localization [17]. [64] and [12] proposed proactive methods to protect images from manipulation based on image watermarking and steganography. Also, reconstruction errors through the inversion process of DDIM [44] are studied by prior works [55, 33, 41, 32] for diffusion generative content detection. Moreover, the previous work [63, 46, 34, 18] develops specific techniques that increase the generalization to unseen forgeries. For example, HiFi-Net [18] proposes a tree structure to model the inherent hierarchical correlation among different forgery methods, NPR [46] devises a special representation as generative artifacts, and the training set diversity can also contribute to generalization ability [34]. Unlike prior works, our MM-Det leverages multi-modal reasoning to achieve a high level of generalization ability.

Video-level Detector Early video-level methods primarily focused on detecting facial forgery. For example, [25] learned the boundary artifacts between the original background and manipulated faces. [19] discriminated fake videos from the inconsistency of mouth motion. [73] designed a multi-attentional detector to capture deepfake features and artifacts. F3Net [38] captured global and local forgery traces in the frequency domain. [10, 15, 74, 56] explored temporal information and inconsistency from fake videos. Most recently, DD-VQA [70] formulates deepfake detection as a sentence-generation problem, largely improving the interpretation of deepfake detection. However, these studies are restricted to facial forgery methods, which are insufficient for the current defensive systems that address diverse content produced by diffusion models. Therefore, we develop MM-Det to detect diffusion video content, pushing forward the frontier of forgery video detection.

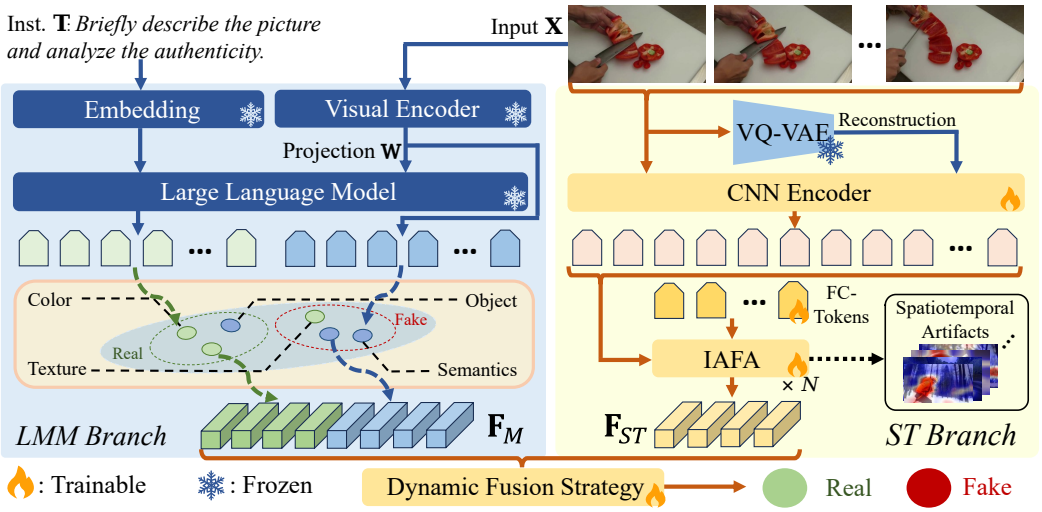


Figure 4: Multi-Modal Detection network (MM-Det) architecture. Given an input video $\mathbf{X}$, the Large Multi-modal Model (LMM) branch takes the frame and instructions to generate Multi-Modal Forgery Representation (MMFR). Hidden states from the visual encoder and large language model are extracted to form the MMFR, denoted as $\mathbf{F}_M$, which helps capture the forgery traces among different diffusion-generated videos. In the Spatio-Temporal (ST) branch, videos are first reconstructed via a VQ-VAE, and then fed into a CNN encoder, followed by In-and-Across Frame Attention (IAFA) modules detailed in Sec. 3.2. IAFA is introduced to capture features based on spatial artifacts and temporal inconsistencies, termed as $\mathbf{F}_{ST}$. At last, a dynamic fusion strategy combines $\mathbf{F}_M$ and $\mathbf{F}_{ST}$ for the final forgery prediction.

Large Multi-modal Models (LMMs) LMMs possess generalizable problem-solving abilities in real-world tasks, including object detection[14], semantic segmentation[67] and visual question answering[59]. [2, 23, 27, 28] studied feature alignment schemes to bridge visual and textual domains for LMMs. [68, 72, 71] extended the boundaries of LMMs to multi-modal downstream tasks. [45] aligns multi-modal features for capabilities on cross-domain behaviors. [58] developed a Large Language Model (LLM)-based feature extractor for cheap-fake detection. Inspired by these studies, we stimulate the powerful perceptual and reasoning ability of an LMM by introducing the multi-modal feature space in video forgery detection.

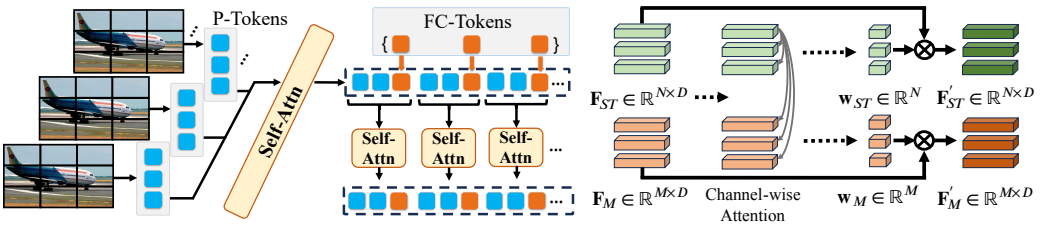
3 Methods

In this section, we introduce the Multi-Modal Detection (MM-Det) framework for diffusion video detection, as depicted in Fig. 4. More formally, Sec. 3.1 details a Large Multi-modal Model (LMM) branch that learns a Multi-Modal Forgery Representation (MMFR). Then Sec. 3.2 reports a Spatio-Temporal (ST) branch that utilizes In-and-Across Frame Attention (IAFA) to capture spatial artifacts and temporal inconsistencies in forged videos. Lastly, a dynamic fusion technique reported in Sec. 3.3 adaptively combines outputs from the LMM branch and ST branch.

3.1 Multi-Modal Forgery Representation

We propose a novel Multi-Modal Forgery Representation (MMFR) from the multi-modal space of LMMs in LMM branch. This representation utilizes the powerful perceptual and reasoning abilities of LMMs in the form of instruction-based conversations.

Specifically, LMM branch is built on the top of LLaVA [28], one representative LMM, which has two key components: a visual encoder (*e.g.*, $\mathcal{D}_v$), instantiated by visual encoders from the Contrastive Language-Image Pre-Training (CLIP) [39], and the large language model $\mathcal{D}_L$ (*i.e.*, Llama 2 [50]). Let us denote the input video as $\mathbf{X} \in \mathbb{R}^{N \times H \times W \times C}$ that contains N frames, where each frame is represented as $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$. First, $\mathbf{x}$ is fed to $\mathcal{D}_v$ to obtain the corresponding visual representation $\mathbf{F}_V \in \mathbb{R}^Z$. Such $\mathbf{F}_V$ not only contains rich semantics but also shows impressive generalization ability and robustness in the forgery detection task [36, 43, 9]. Then, a textual instruction $\mathbf{T}$ is sampled from



(a) Mechanism of In-and-Across Frame Attention (b) Dynamic Fusion Strategy

Figure 5: (a) In-and-Across Frame Attention (IAFA): Each input frame (or its feature map) is divided into patches that are transformed into tokens, termed P-tokens (■). We introduce additional frame-centric tokens FC-tokens (■) encapsulating the global forgery information of the video frame. In each transformer layer, self-attention is applied alternately among all P-tokens from video frames as well as among the same frame's P-tokens and its FC-tokens. (b) The dynamic fusion strategy captures that takes $\mathbf{F}_{ST}$ and $\mathbf{F}_M$ as inputs and output channel-wise dependencies, which help refine forgery representations for the fusion.

pre-defined templates $\mathbf{Q}$ to guide the LMM on forgery detection reasoning. Both visual representation $\mathbf{F}_V$ and instruction $\mathbf{T}$ are fed to $\mathcal{D}_L$, which generates enhanced visual representations (*e.g.*, $\mathbf{F}_L$). We convert $\mathbf{F}_V$ into a sequence of visual tokens [28] (*i.e.*, $\mathbf{H}_v = \{\mathbf{h}_{v,m}\}_{m=1}^{M} \in \mathbb{R}^{M\times D}$), and $\mathbf{T}$ is transformed into textual tokens $\mathbf{H}_t = \{\mathbf{h}_{t,o}\}_{o=1}^{O} \in \mathbb{R}^{O\times D}$. Both $\mathbf{H}_v$ and $\mathbf{H}_t$ are taken as the input to $\mathcal{D}_L$, generating $\mathbf{F}_L \in \mathbb{R}^{S\times D}$ that can be tokenized into the language response providing reasoning (Fig. 2) about the authenticity of the input $\mathbf{x}$. This procedure is formulated as

$$\mathbf{F}_L = \mathcal{D}_L(\mathbf{H}_t, \mathbf{H}_v) = \mathcal{D}_L(\mathbf{T}, \mathbf{F}_V), \tag{1}$$

where $\mathbf{T}$ guides the pre-trained $\mathcal{D}_L$ in comprehending visual content (*i.e.*, $\mathbf{F}_V$), discerning the subset information from $\mathbf{F}_V$. This instruction $\mathbf{T}$ enables LMM branch to obtain the multi-modal representation that leverages the generalization ability from the pre-trained large language model Llama 2 (*i.e.*, $\mathbf{D}_L$), being different to prior work [36, 9] that only relies on $\mathbf{F}_E$.

Lastly, we retrieve the final MMFR, denoted as $\mathbf{F}_M \in \mathbb{R}^{M\times Z}$, by concatenating $\mathbf{F}_V$ and $\mathbf{F}_L$ after a linear layer(*i.e.*, `PROJ`), as

$$\mathbf{F}_M = \texttt{CONCAT}(\{\texttt{PROJ}(\mathbf{F}_V), \texttt{PROJ}(\mathbf{F}_L)\}). \tag{2}$$

3.2 Capturing Spatial-Temporal Forgery Traces

Targeting capturing spatiotemporal artifacts in video tasks, we introduce a Spatial-Temporal (ST) branch that learns effective diffusion forgery representation at the video level. Through a reconstruction procedure, we amplify the diffusion traces in the frequency domain, which is then captured by In-and-Across Frame Attention (IAFA) to form an effective video-level feature.

Amplification of Diffusion Traces Similar to prior studies [55, 33] that discovered specific generative traces of diffusion models through reconstruction on diffusion-generated images, we utilize an Autoencoder to amplify diffusion traces in videos. The reconstruction procedure can be expressed as follows.

More formally, denote the input video as $\mathbf{X} \in \mathbb{R}^{N\times H\times W\times C}$ that contains N frames, in which each frame is represented as $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$. We leverage a VQ-VAE [51] to obtain the reconstructed version of $\mathbf{x}$, which is denoted as $\hat{\mathbf{x}} \in \mathbb{R}^{H\times W\times C}$. The difference between $\mathbf{x}$ and $\hat{\mathbf{x}}$ makes an effective indicator of showing if the input is generated by diffusion models, as depicted in Fig. 3. Therefore, we jointly proceed $\mathbf{x}$ and $\hat{\mathbf{x}}$ to the following proposed modules for learning an effective representation of discerning forgeries.

It is worth mentioning that the prior approach [55] also adopts the idea of using the residual difference between original and reconstructed inputs to help forgery detection, but the reconstruction method requires multiple time-step denoising operations, which are computationally infeasible to reconstruct all frames from $\mathbf{X}$. In contrast, our VQ-VAE-based reconstruction method only requires one single forward propagation to obtain reconstructed frame $\hat{\mathbf{x}}$, meanwhile preserving the effectiveness in indicating the discrepancy between real and fake inputs.

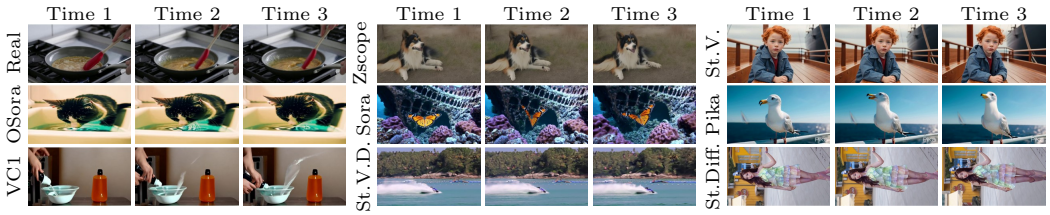


Figure 6: Sampled videos from DVF dataset. DVF contains 8 video generation methods, including 7 text-to-video methods and 1 image-to-video method. Real videos are selected from Internvid-10M [54] and Youtube-8M [1]. [Key: OSora: OpenSora; VC1: Videocrafter1 [5]; Zscope: Zeroscope; St. V. D.: Stable Video Diffusion [4]; St.Diff.: Stable Diffusion [42]; St. V.:Stable Video]

Integration of Spatial and Temporal Information Same as the previous work [3, 35] that utilizes ViT to learn video-level information, we transform each input video frame (*i.e.*, $\mathbf{x}$) into L tokens, and propose IAFA for information aggregation, as depicted in Fig. 5a. Let us denote frame-level tokens as Patch-wise tokens (P-tokens), as they represent local information of one patch from $\mathbf{x}$. More formally, the ith frame, $\mathbf{x}_i$ is divided into L patches, and all patches are projected into $\mathbf{T}_i = \{\mathbf{t}_i^j\}_{j=1}^{L} \in \mathbb{R}^{L\times D}$, where D represents the dimension of each P-token. Also, to capture the global forgery information for each video frame, we introduce additional tokens called Frame-Centric tokens (FC-tokens). The FC-token is denoted as $\mathbf{p}_i \in \mathcal{R}^D$ for frame $\mathbf{x}_i$ and attends other tokens *within* the same frame $\mathbf{x}_i$.

During the forward propagation, we conduct IAFA based on a Transformer, with each block containing two self-attentions and consecutively modeling the local and global forgeries at each video frame. Specifically, the first self-attention captures dependencies among P-tokens that restore local forgery clues. This is formulated by Eq. 3 that $\mathbf{t}_i^j \in \mathbb{R}^D$ attends the token $\mathbf{t}_q^p \in \mathbb{R}^D$ that represents pth token from qth frame $\mathbf{x}_q$. Consequently, given the ith frame *i.e.*, $\mathbf{x}_i$, the second self-attention is conducted among the FC-token (*i.e.*, $\{\mathbf{p}_i\}$) and P-tokens (*i.e.*, $\{\mathbf{t}_i^0, \mathbf{t}_i^1, \ldots \mathbf{t}_i^{L-1}\}$) from the same frame, which encapsulates patch-wise forgery information into the global one for learning the more robust representation. We formulate this procedure in Eq. 4.

$$\mathbf{t}_i^j = \sum \texttt{ATTN}(\mathbf{t}_i^j, \mathbf{t}_q^p) \quad i,j \in [1,N], j,p \in [1,L], \tag{3}$$

$$\mathbf{p}_i = \sum \texttt{ATTN}(\mathbf{p}_i, \mathbf{t}_i^j) \quad j \in [1,L], \tag{4}$$

where ATTN refers to the self-attention operation.

3.3 Dynamic Fusion

We devise the dynamic fusion strategy (*i.e.*, $\mathcal{D}_f$) that combines spatiotemporal information from ST branch and MMFR (*i.e.*, $\mathbf{F}_f$ and $\mathbf{F}_m$) for the final prediction, by adjusting their contributions based on forgeries from the input. More formally, $\mathcal{D}_f$ (Fig. 5b) learns channel-wise dependencies among forgery representations (*e.g.*, $\mathbf{F}_{ST}$ and $\mathbf{F}_M$) via the attention mechanism, generating $\mathbf{w} \in \mathbb{R}^{N+M}$ as the output. This procedure can be expressed as $\mathbf{w} = \mathcal{D}_f(\texttt{CONCAT}\{\mathbf{F}_{ST}, \mathbf{F}_M\})$. Also, $\mathbf{w}$ contains $\mathbf{w}_{ST} \in \mathbb{R}^N$ and $\mathbf{w}_M \in \mathbb{R}^M$, representing learned channel-wise weights for $\mathbf{F}_{ST}$ and $\mathbf{F}_M$, respectively. Such channel-wise weights are important in the fusion purpose, as they help emphasize useful information — we use $\mathbf{w}_{ST}$ and $\mathbf{w}_M$ to refine forgery representations as $\mathbf{F}'_{ST} = \mathbf{F}_{ST}\mathbf{w}_{ST}$ and $\mathbf{F}'_M = \mathbf{F}_M\mathbf{w}_M$. Lastly, $\mathbf{F}'_{ST}$ and $\mathbf{F}'_M$ are concatenated into the fused representation $\mathbf{F}_0 \in \mathbb{R}^{(M+N)\times D}$, which is used for the final scalar prediction s via the average pooling (*i.e.*, AVG) and linear layers(*i.e.*, PROJ):

$$s = \texttt{PROJ}(\texttt{AVG}(\mathbf{F}_0)) = \texttt{PROJ}(\texttt{AVG}(\texttt{CONCAT}\{\mathbf{F}'_{ST}, \mathbf{F}'_M\}). \tag{5}$$

4 Diffusion Video Forensics (DVF) Dataset

We construct a large-scale dataset for the video forensic task named Diffusion Video Forensics (DVF), as shown in Fig. 6. DVF contains 8 diffusion generative methods, including Stable Diffusion [42], VideoCrafter1 [5], Zeroscope, Sora, Pika, OpenSora, Stable Video, and Stable Video Diffusion[4].

To efficiently streamline the collection, we construct an effective automated pipeline that generates forgery videos based on real videos and prompts. Specifically, we start from two real video datasets,

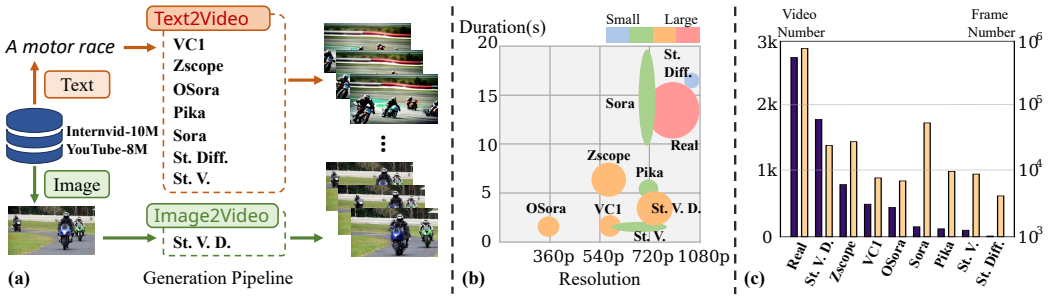


Figure 7: The overview of DVF dataset: (a) The procedure of forged video generation and collection. Real frames and captions are sampled from Internvid-10M [54] and Youtube-8M [1] for text-to-video and image-to-video generation, respectively. (b) DVF contains videos at various resolutions and durations. (c) The scale of each video dataset in DVF is measured by the frame and video numbers.[Key: VC1: Videocrafter1 [5]; Zscope: Zeroscope; OSora: OpenSora; St.Diff.: Stable Diffusion [42]; St. V.:Stable Video; St. V. D.: Stable Video Diffusion [4]]

Internvid-10M [54] and Youtube-8M [1]. Real videos are sampled for rich semantic content, with their frames and captions used for generation. Fig. 7 a introduces the generation process of DVF. For open-sourced generation methods, a prompt is fed to a text-to-video method(*i.e.* VideoCrafter1, Zeroscope, OpenSora), or a frame is provided to an image-to-video method(Stable Video Diffusion) to generate the corresponding fake video. For commercial and close-sourced datasets(*i.e.* Stable Diffusion, Stable Video, Sora, Pika), forgery videos are collected from official websites and social media. In total, we collect $3,938$ fake videos and $2,750$ real videos in DVF. As shown in Fig. 7 b and 7 c, our dataset contains multiple resolutions and durations. The video number of each dataset varies from 0.1k to 2.8k, with the corresponding frame numbers from 4.2k to 784k. More details about DVF are provided in Appendix A.3.1.

5 Training Strategy

This section details our two-stage training strategy, in which we first finetune the LMM branch via instruction tuning [28] and then optimize the entire framework in an end-to-end manner.

LMM Branch Instruction Tuning We first adapt LLaVA [28] to the forgery detection downstream task based on instruction tuning, an empirically effective way for various downstream tasks, which leverages LoRA [21] to improve the reasoning ability of Large Language Models(LLMs). For that, construct a large image-text paired dataset, named Rich Forgery Reasoning Dataset. Please refer to Appendix A.3.2 for more details. We use multi-turn conversations to fine-tune the LMM, enhancing its ability to identify and judge the authenticity of input images. Following the instruction tuning strategy of LLaVA [28], we only fine-tune the projection layers and LLM in LLaVA. More formally, we formulate the objective function as the loss for an auto-regressive model, which is based on answer tokens from the LLM, as:

$$\mathcal{L}(\theta_1) = -\sum_{t=1}^{T} log(p_{\theta_1}(s^t|s^{i<t})), \tag{6}$$

where s^i refers to the i^{th} prediction token, T refers to the length of total prediction tokens, and θ_1 refers to the trainable parameters in the LMM.

End-to-End Training After fine-tuning LLaVA, we use this model to form LMM branch of MM-Det, and then the entire model is trained in an end-to-end manner. Please note that all parameters in LMM branch are frozen to ensure the optimal multi-modal representation can be obtained. More formally, we denote MM-Det's final prediction scalar and the ground truth as s and y, respectively, and the model is optimized by the cross-entropy loss $\mathcal{L}$ as follows:

$$\mathcal{L}(\theta_2) = -(y \log \mathcal{D}(v) + (1-y) \log(1-\mathcal{D}(v)) \tag{7}$$

where θ_2 refers to trainable parameters in both ST branch and dynamic fusion modules.

Table 1: Video forgery detection performance on the DVF dataset measured by AUC (%). [Key: **Best**; Second Best; Stable Diff.: Stable Diffusion; Avg.: Average]

Method	Video-Crafter1	Zero-scope	Open-Sora	Sora	Pika	Stable Diff.	Stable Video	Avg.
CNNDet [53]	83.3	70.2	81.9	63.8	76.5	71.8	80.8	75.5
DIRE [55]	56.8	61.9	56.1	60.7	70.0	58.3	71.2	62.1
Raising [8]	63.9	58.5	64.6	62.4	66.0	91.3	59.5	66.6
Uni-FD [36]	93.6	90.1	83.9	85.4	93.0	81.5	87.9	87.9
F3Net [38]	96.1	91.8	85.9	66.0	95.6	86.3	**96.0**	88.2
ViViT [3]	89.2	88.0	85.2	81.6	92.7	88.1	92.1	88.1
TALL [62]	76.5	61.8	69.8	62.3	79.9	85.9	64.8	71.6
TS2-Net [31]	60.7	72.0	74.3	81.0	80.2	60.2	80.2	72.7
DE-FAKE [43]	72.3	70.3	53.6	67.3	88.4	86.0	74.1	73.1
HiFi-Net [18]	96.7	93.9	94.9	83.9	85.8	80.2	87.3	89.0
MM-Det (Ours)	**97.4**	**98.6**	**97.6**	**91.7**	**98.0**	**92.1**	95.1	**95.7**

6 Experiments

6.1 Setup

In the experiment, we use the proposed DVF for the evaluation. In training, $1,000$ videos from YouTube and $1,800$ fake videos generated by Stable Video Diffusion serve as the training set, in which 80% are used for training and the remaining 20% for validation. Real videos from Internvid-10M [54] and fake videos from 6 generative methods are used as testing samples. More details on training and testing are provided in Appendix A.4.

For a fair comparison, we choose the following 10 recent detection methods as baselines. CNNDet [53] applies a ResNet [20] as the backbone for forgery detection. F3Net [38] utilizes frequency traces left in forgery content. HiFi-Net [18] devise a specific hierarchical fine-grained learning scheme to learn a wide range of forgery traces. Clip-Raising [9], Uni-FD [36] takes advantage of a pre-trained CLIP [39] as a training-free feature space. DIRE [55] detects diffusion images based on a reconstruction process of DDIM [44]. ViViT [3], TALL [62], and TS2-Net [31] take advantage of spatiotemporal information in various visual tasks. DE-FAKE [43] adopts visual and textual representations based on a CLIP encoder for image forgery detection. For the measurement, we choose AUC since it is a threshold-independent metric.

6.2 Video Forgery Detection Performance

In Tab. 1, our proposed MM-Det achieves SoTA performance in detecting diffusion video, surpassing the second-best method, *i.e.*, HiFi-Net, by 6.7% in AUC scores. Specifically, for prior methods that are based on pre-trained CLIP features, such as Raising [8] and Universal FD [36], they remain effective on certain types of diffusion content (*i.e.*, Sora, Pika, Stable Diffusion), but fail on most others. Simple structures like CNN [53] exceed these CLIP-based methods after being fine-tuned on our proposed DVF, which proves the necessity of such datasets. As for our method, MM-Det outperforms other methods in most datasets. Compared with frequency-based forgery methods, *e.g.*, F3Net [38] and CLIP-based methods [8, 36], our method improves the performance from $+0.7\%$(VideoCrafter1) to $+6.3\%$(Sora). It is worth mentioning that HiFi-Net makes the second-best performer in our DVF dataset, achieving 89.0% AUC scores. We believe this indicates the multi-branch feature extractor used in HiFi-Net carries versatile forgery traces at multiple resolutions, enhancing the learning of the forgery invariant. The failure of frequency traces and CLIP features raises the need for more effective features. As for spatiotemporal baselines [3, 62, 31], we outperform them by $+7.6\%$(ViViT), $+24.1\%$(TALL) and $+23.0\%$(TS2-Net), demonstrating the effective features of MMFR and IAFA. At last, our detector improves by 22.6% to another multi-modal detector [57], which utilizes visual information and corresponding captions for feature enhancement. It is shown that the introduction of MMFR is more generalizable than a simple combination of visual features and text descriptions in that the powerful perceptual and reasoning abilities of LMMs play a crucial role in discriminating between real and fake content.

Table 2: Ablation analysis measured by AUC (%). [Key: **Best**; Avg.: Average; Rec.: Diffusion Reconstruction Procedure; Fus. Dynamic Fusion Strategy].

Modules					Video-Crafter1	Zero scope	Open Sora	Sora	Pika	Stable Diff.	Stable Video	Avg.
ViT	Rec.	IAFA	MMFR	Fus.								
✔					69.7	84.8	66.5	54.5	80.6	91.8	87.6	76.5
✔	✔				72.4	77.1	72.7	65.9	83.6	84.1	87.1	77.6
✔	✔	✔			94.4	94.2	82.0	82.0	95.4	**92.1**	93.9	90.6
			✔		90.5	89.1	91.8	86.0	94.2	90.1	92.1	90.5
✔	✔	✔	✔		94.8	94.2	93.2	90.9	94.9	90.6	**97.4**	93.7
✔	✔	✔	✔	✔	**97.4**	**98.6**	**97.6**	**91.7**	**98.0**	**92.1**	95.1	**95.7**

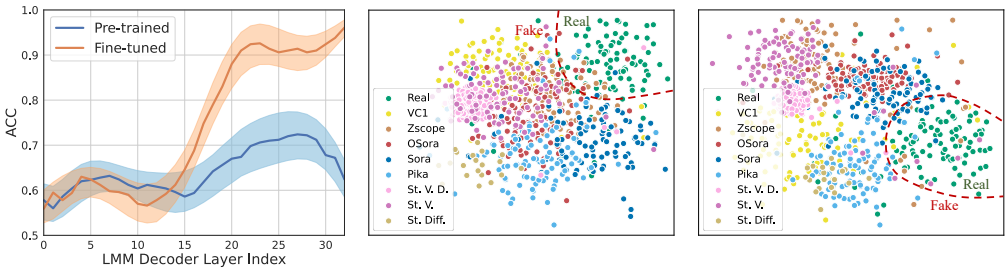
Figure 8: Visualization of artifacts captured from our IAFA and ViViT [3]. We use activation maps to highlight spatial weights within each frame. All content is generated by VideoCrafter1. Features from the last layer of transformers are extracted for visualization.

6.3 Ablation Study

Tab. 2 shows the impact of individual modules proposed in MM-Det. Specifically, we use the Hybrid ViT [11] as the base model and incorporate it with the reconstruction procedure for diffusion trace amplification, which enhances the detection performance by $+1.1\%$ AUC score. Such a module raises the performance in OpenSora and Sora, revealing that frequency-based information benefits forgery detection on these methods. Detection performance is further increased by using IAFA, which strengthens the learning between in-frame and cross-frame information, increasing a $+13.0\%$ AUC score to the base model. The rise in performance indicates such temporal information benefits most types of forgery video detection. After that, Tab. 2's line 4 indicates the effectiveness of MMFR: a detector purely based on such representation can receive 90.5% performance in AUC, $+14.0\%$ higher than the base model. In addition, by merging MMFR (*i.e.*, LMM branch) and ST branch, the performance rises by $+3.2\%$. Finally, with the dynamic fusion strategy, our method receives an impressive 95.7% AUC score for all generative methods, higher than every single feature. These experiments highlight the necessities of each module in our framework. Moreover, an ablation study on LLMs is detailed in Appendix A.5 to prove the effectiveness of various LLMs in MM-Det.

6.4 Spatio Temporal Information Anaylsis

In the analysis of IAFA, we visualize feature activation maps from the last layer of ViT in the ST branch based on the L2-norm. We compare our feature maps with another spatiotemporal baseline, ViViT [3], as depicted in Fig. 8. While attention maps of ViViT are sparse and irregular, the ones of our IAFA have a tendency to concentrate on the segmentation of diffusion-generated objects, indicating that IAFA captures typical spatial forgery regions in frames. The attention mainly focuses on common forgery traces, such as blurred generative patterns and defective parts of objects, signaling that diffusion models might find it difficult to generate delicate content. The concentration of activation on certain informative objects discloses both spatial artifacts of existing generative methods, demonstrating the effectiveness of our proposed ST branch.



(a) (b) (c)

Figure 9: (a): Clustering accuracy using features from different layers in LMM branch showcases MMFR's ability to discern forgeries. (b)(c): t-SNE [52] visualization of features from ST and LMM branches. For each dataset, 100 videos are sampled and clustered for good visibility. Both features demonstrate boundaries between real and forgery videos. (b) Features from the ST branch. (c): Features from the LMM branch.

6.5 Multimodal Forgery Representation Analysis

Fig. 9 details the effectiveness of MMFR in LMM branch. First, as depicted in Fig. 9 a, we quantify the detection ability of features from each Transformer-based decoder layer in the LLM. Specifically, we evaluate both pre-trained LLaVA [28] and its fine-tuned version on the task of distinguishing real and fake frames. These frames are randomly sampled from $1,000$ real videos and equivalent fake ones from Stable Video Diffusion [42]. Layer-wise outputs from the large language model in LLaVA (*i.e.*, Vicuna [50]) are obtained — for i th layer in the LLM, we denote its output features as $\mathbf{f}_i^o$, $i \in [1, 33]$. The K-Means clustering algorithm is adopted to evaluate the classification accuracy based on $\mathbf{f}_i^o$. Empirically, we observe that features extracted from a fine-tuned LLaVA show promising classification accuracy in the last few layers, *e.g.*, 22-nd or later layers. This phenomenon indicates that specific layers in LLMs indeed generate features that can be used for image forensic tasks. Such features are utilized in MMFR to exhibit high generalization ability towards diverse and unseen forgeries. Secondly, the comparison between pre-trained and fine-tuned LLaVA highlights the importance of downstream task-oriented instruction tuning for LMMs. This conclusion is consistent with the findings of prior works [27, 28, 69, 68].

In addition, shown in Fig. 9 b and 9 c, we analyze features from ST branch and LMM branch through t-SNE [52]. Both features achieve superior performance in separating real and forgery videos. In Fig. 9 b, spatiotemporal information forms a rough boundary between real and fake videos. This feature is effective for VideoCrafter1 [5], Zeroscope, Stable Video, and Pika, whose durations and resolutions are similar to the training set. However, the detection performance might decrease on Sora and OpenSora with overlay in the clusters. We suppose that various resolutions and durations may compromise the generalization ability, magnifying the importance of a comprehensive dataset for these videos. Fig. 9 c demonstrates the more powerful feature from LMM branch. Samples from Zeroscope, Sora, and Pika are compacted into a denser area, indicating the ability of LLMs to conduct generalizable reasoning. Such features provide new insights for detection when spatial and temporal artifacts are not obvious among the latest forgery videos.

7 Conclusion

In this work, we develop an effective video-level algorithm termed Multi-Modal Detection (MM-Det) for diffusion-generated video detection. MM-Det leverages a novel generalizable Multi-Modal Forgery Representation (MMFR) that is obtained from multi-modal spaces in LMMs. Specifically, the proposed MM-Det has two major branches: the LMM branch, which incorporates vision and text features from the fine-tuned foundation model, and the ST branch, which concentrates on modeling spatial-temporal information aggregated through In-and-Across Frame Attention. Extensive experiments demonstrate the effectiveness of our proposed detector. In addition, we establish a comprehensive dataset for various diffusion generative videos, which we hope will serve as a general benchmark for real-world video forensic tasks.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China under Grant 62301310 and 62225112, and in part by Sichuan Science and Technology Program under Grant 2024NSFSC1426.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[6] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 973–982, 2023.

[7] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[8] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. *arXiv preprint arXiv:2312.00195*, 2023.

[9] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024.

[10] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Kang Fu, Xiaohong Liu, Jun Jia, Zicheng Zhang, Yicong Peng, and Jia Wang. Rawiw: Raw image watermarking robust to isp pipeline. *Displays*, 82:102637, 2024.

[13] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021.

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[15] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021.

[16] Xiao Guo, Vishal Asnani, Sijia Liu, and Xiaoming Liu. Tracing hyperparameter dependencies for model parsing via learnable graph pooling network. In *Proceeding of Thirty-eighth Conference on Neural Information Processing Systems*, Vancouver, Canada, December 2024.

[17] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. In *International Journal of Computer Vision*, December 2024.

[18] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[19] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[25] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[26] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. *arXiv preprint arXiv:2403.14333*, 2024.

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[29] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[30] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

[31] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer, 2022.

[32] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024.

[33] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023.

[34] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022.

[35] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021.

[36] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[37] Yongyang Pan, Xiaohong Liu, Siqi Luo, Yi Xin, Xiao Guo, Xiaoming Liu, Xiongkuo Min, and Guangtao Zhai. Towards effective user attribution for latent diffusion models via watermark-informed blending. *arXiv preprint arXiv:2409.10958*, 2024.

[38] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[40] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

[41] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. *arXiv preprint arXiv:2401.17879*, 2024.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[43] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[45] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

[46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[47] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[49] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[50] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[51] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[53] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[54] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[55] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[56] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023.

[57] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

[58] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. Cheap-fake detection with llm using prompt engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 105–109. IEEE, 2023.

[59] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[60] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9543–9552, 2019.

[61] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[62] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023.

[63] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.

[64] Yiwei Yang, Zheyuan Liu, Jun Jia, Zhongpai Gao, Yunhao Li, Wei Sun, Xiaohong Liu, and Guangtao Zhai. Diffstega: Towards universal training-free coverless image steganography with diffusion models. *arXiv preprint arXiv:2407.10459*, 2024.

[65] Yuguang Yao, Xiao Guo, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu, Sijia Liu, et al. Reverse engineering of deceptions on machine-and human-centric attacks. *Foundations and Trends® in Privacy and Security*, 6(2):53–152, 2024.

[66] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.

[67] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[68] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[69] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[70] Yue Zhang, Ben Colman, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deep fake detection. *arXiv preprint arXiv:2402.00126*, 2024.

[71] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

[72] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*, 2024.

[73] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[74] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021.

[75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix / Supplemental Material

A.1 Limitations

Although our proposed MM-Det advances in detecting fake videos, further issues are left to handle. First, as the landscape of video manipulation technology evolves, new techniques and tools will outpace existing detection methods. The gap between training data and real-world applications can lead to misleading results. Currently, our method is limited to fully synthesized diffusion videos, lacking the generalization to more delicate forgeries like partial manipulation. A possible reason is that small forgery traces disappear after multiple downsampling operations in the deep network of LMMs. Besides, the integration of a large language model into detection costs huge computational complexity, which is not an optimal choice for an environment with limited resources.

In conclusion, while our algorithm makes a critical step forward in detecting fake videos, it faces significant challenges due to the rapid advancement of video manipulation technologies. Addressing these limitations requires further research to keep pace with the evolving techniques in digital content manipulation.

A.2 Broader Impacts

In this work, our team has developed an effective algorithm to detect fake videos, a breakthrough that promises to fortify the authenticity of online media. In real-world social media where misinformation can spread rapidly, our method acts as a crucial safeguard by empowering platforms to flag and remove deceptive videos before they can mislead users. However, our methods may fail in extreme situations, such as blurred images or noisy images. The algorithm should be carefully treated to avoid misleading results. The long-term influence of our work protects public trust by ensuring the authenticity of digital content. Meanwhile, precaution is needed for a fair application.

A.3 Datasets

A.3.1 Diffusion Video Forensics

We propose a comprehensive dataset, named Diffusion Video Forensics (DVF), for diffusion video forensics, as shown in Tab. S1. DVF consists of fake videos generated from 8 different generation methods, covering text-to-video and image-to-video generative methods. In total, We make a collection of $2,750$ real videos and $3,938$ fake videos. Real videos are from YouTube and Internvid-10M [54].

We formally introduce the generation pipeline for video collection. Generation methods in DVF are divided into closed-sourced methods and open-sourced methods. For closed-sourced methods(Sora, Pika, Stable Diffusion [42] and Stable Video), we collect video samples from official websites and social media like TikTok to form the forgery video datasets. For open-sourced methods, the generation pipelines are divided into text-to-video (OpenSora, VideoCrafter1 [5] and Zeroscope) and image-to-video (Stable Video Diffusion [4]). For a text-to-video generative method, real data derives from a text-image paired video dataset, Internvid-10M. Specifically, we fetch paired real videos $R = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\}$ and corresponding captions $C = \{c_1, c_2, ..., c_N\}$. We directly apply the

Table S1: Diffusion Video Forensics Composition [Key: T2V: Text-to-Video; I2V: Image-to-Video]

Dataset	Source	Video Number	Resolution
Real	Youtube & Internvid-10M	2750	1280×720
Stable Video Diffusion	I2V	1800	1024×576
VideoCrafter1	T2V	450	1024×576
Zeroscope	T2V	800	1024×576
Sora	Social Media	153	1280×720
Pika	T2V, I2V	122	1280×720
OpenSora	T2V	500	512×512
Stable Diffusion	I2V	12	1080×1920
Stable Video	T2V	101	1024×576, 1920×1080

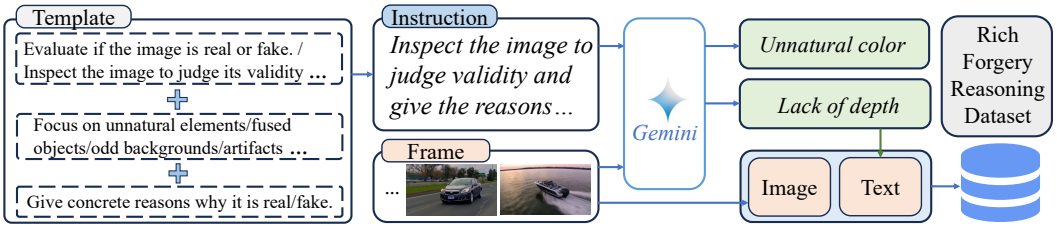


Figure S1: The overview of Rich Forgery Reasoning Dataset. To obtain text-image paired data on forgery reasoning, we take advantage of the powerful reasoning ability of Gemini [48] to generate ground truth for each image. First, an instruction T is sampled from a template, along with a frame $\mathbf{f} \in \mathbb{R}^{H\times W\times C}$ fed into Gemini for forgery analysis and detection. The response r contains detailed judgment and reasoning on the content and authenticity of $\mathbf{f}$(*e.g.*, analyses on color and depth). Finally, $\mathbf{f}$ and r are collected as the text-image paired data to form Rich Forgery Reasoning Dataset.

captions as prompts to generate fake video datasets F, such that $F = \{\mathbf{f}_1, \mathbf{f}_2, ...\mathbf{f}_N\}, \mathbf{f}_i = g_t(c_i), i \in [1, N]$, where g_t denotes a text-to-video method. Both R and F are included in DVF as real and fake datasets. For image-to-video methods, real videos come from Youtube-8M [1], which are denoted as $R = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\}, \mathbf{r}_i \in \mathbb{R}^{L\times H\times W\times C}$. For each video r_i, a real frame $x_i \in \mathbb{R}^{H\times W\times C}$ is randomly sampled from r_i and serves as the conditional input for generation. The fake datasets are obtained as $F = \{\mathbf{f}_1, \mathbf{f}_2, ...\mathbf{f}_N\}, \mathbf{f}_i = g_{im}(r_i), i \in [1, N]$, where g_{im} denotes an image-to-video method.

A.3.2 Rich Forgery Reasoning Dataset

We construct a text-image paired dataset, called Rich Forgery Reasoning Dataset (RFRD), to support instruction-tuning LMMs on the forgery detection task, as shown in Fig. S1. We start from the YouTube and Stable Video Diffusion dataset in DVF, where we select $1,000$ real videos and $1,800$ fake videos. Depending on the powerful reasoning ability of Gemini [48] v1.5 Pro, we follow the scheme in Fig. S1 to generate ground truth for frames. In total, $1,921$ real frames and $3,579$ fake frames are sampled to generate $5,500$ image-text paired textual descriptions. These descriptions are then cleaned and converted into 38k multi-turn conversations for fine-tuning LLaVA [28] in LMM branch of MM-Det.

A.4 Implementation Details

Hyperparameters of MM-Det We introduce the implementation of our MM-Det. In ST branch, we employ a Hybrid-ViT [11] to build the video-level feature encoder. We choose Hybrid-ViT-B, with a ViT-B/16 on top of a ResNet-50 backbone, the patch size 14×14, and the hidden size 768. We employ IAFA in each attention block of the ViT, with FC-tokens initialized as the class token of ViT. In the embedding stage, learnable spatial and temporal embeddings are introduced in the form of addition. All patches at the same position within frames share the same spatial embedding and patches with the same timestep share the same time embedding. A pre-trained VQ-VAE is applied to reconstruct videos, with hidden size $d = 256$ and the codebook size $K = 512$. We train the VQ-VAE on $50,000$ images from ImageNet. In LMM branch, we utilize LLaVA [28] v1.5, with a CLIP [39] encoder $\mathcal{E}$ of CLIP-ViT-L-patch14-336 and a large language model $\mathcal{D}$ of Vicuna-7b for reasoning. Our proposed MMFR is composed of the pooler output $\mathbf{F}_V \in \mathbb{R}^{1024}$ from the last layer of $\mathcal{E}$ and the embedding of output $\mathbf{F}_L$ from the last layer of $\mathcal{D}$. To balance effectiveness and efficiency, we fix the length of output tokens into 64 and obtain $\mathbf{F}_L \in \mathbb{R}^{64\times 4096}$.

Training and Inference As for the experimental resources in training and inference, we conduct all experiments using a single NVIDIA RTX 3090 GPU and a maximum of 200G memory.

The training strategy of MM-Det is in two-stage. We first conduct instruction tuning for LLaVA in the LMM branch based on LoRA [21]. We start from a pre-trained LLaVA v1.5 and train it on our collected Rich Forgery Reasoning Dataset detailed in Sec. A.3.2. We use an Adam optimizer with the learning rate set as $2e^{-5}$ for 10 epochs. After that, we integrate LLaVA into MM-Det and conduct the overall training. The training set is split into $8:2$ for training and validation data. For each video,

Table S2: Evaluation on different Large Language Models measured by AUC(%). [Key: **Best**; Avg.: Average].

LLM	Video-Crafter1	Zeroscope	OpenSora	Sora	Pika	Stable Diffusion	Stable Video	Avg.
N/A	94.4	94.2	82.0	82.0	95.4	92.1	93.9	90.6
Vicuna-7b	97.4	**98.6**	**97.6**	91.7	98.0	92.1	95.1	95.7
Vicuna-13b	**98.0**	96.2	94.8	**94.4**	**98.6**	**95.2**	**97.0**	**96.3**
Mistral-7b	95.1	96.3	92.6	92.1	95.8	92.9	**97.0**	94.5

Table S3: Perfomance of MM-Det on common post-processing operations measured by AUC(%).

N/A	Blur $\sigma = 3$	JPEG $Q = 50$	Resize 0.7	Rotate 90	Mixed
95.7	89.2	93.2	91.7	92.1	91.9

successive 10 frames are randomly sampled and cropped into 224×224 as the input. We use an Adam optimizer with the learning rate set as $1e^{-4}$ for training until the model converges.

For inference, we evaluate all models at the video level. For frame-level baselines, the score of an entire video is obtained as the average score of all frames. For video-level methods, successive clips are fed according to the corresponding window size, and the entire score is obtained as the average score of all clips. In addition, to leverage the efficiency for inference, MM-Det first caches MMFR for each video by conducting reasoning at the interval of 1 frame every 6 seconds. During inference, each video clip directly applies MMFR from the nearest cached frame as an approximation to reduce the huge computational cost.

A.5 Ablation Study on LLMs

We adopt alternative LLMs in our MM-Det to evaluate different choices of language backbones. Performance is reported in Tab. S2. More formally, for a fair comparison, when using different LLMs, we maintain other components *e.g.*, CLIP, the reconstruction procedure, IAFA, and the dynamic fusion of the original MM-Det remained. Specifically, the introduction of a combined vision and text space from Vicuna-7b in LLaVA improves the performance by $+5.1\%$. As for the choice of LMMs, Vicuna-7b achieves a total 95.7% performance, $+1.2\%$ higher than Mistral-7b. We suppose this result may be attributed to different attention mechanisms in Vicuna and Mistral. Vicuna-13b gains a further improvement by $+0.6\%$ due to incremental parameters in capturing more effective multi-modal feature spaces. These results prove that our MMFR is effective and extensible to other language models.

A.6 Robustness Analysis

To analyze the robustness of our method, we conduct an additional evaluation of MM-Det based on common post-processing operations. We choose Gaussian blur with $\sigma = 3$, JPEG compression with quality $Q = 90$, resize with a ratio of 0.7, rotation with an angle of 90, and a mixture of all operations as unseen perturbations in real-world scenarios. Testing samples are selected from DVF to form a total of 500 real videos and 500 fake videos. As reported in Tab. S3, MM-Det meets a degradation of 2.5%(JPEG Compression) to 6.5%(Gaussian blur), with all performance above 89%. The results indicate the effectiveness of our method under these operations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims fit with the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete and correct proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Although no link is available for our work now, we are strongly intended to share our work in public when the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative social impacts in Boarder Impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper has no risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All creators and owners of assets are properly credited, and the license and terms of use explicitly mentioned and properly are respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented, with the documentation provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper does not involve study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.