PROMPT-DECOUPLED AI-GENERATED MULTIMODAL VIDEO DETECTION AND ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI-generated videos become increasingly realistic, deepfake video detection is vital for safeguarding cybersecurity and video source attribution helps ensure regulatory accountability. Existing deepfake detection methods focus on audio and vision modalities (*e.g.*, the mismatch between lip movements and audio) while largely ignoring the textual modality. Meanwhile, research on video attribution is limited, especially from a multimodal perspective. To address these gaps, we propose a Prompt-Decoupled MultiModal Forensics (PDMF) that combines vision and text modalities for concurrent AI-generated video detection and attribution. PDMF converts video frames into descriptive text subtitles, extracts visual and textual features through advanced encoders, and projects textual features into visual space. By applying the cross-modal residual subtraction, we filter out shared semantic context and successfully expose model-specific artifact features. These features are stored in an extensible reference database to convert video attributes into instance retrieval tasks. By using only 100 videos per class and 2-10 times less time than alternative methods, PDMF achieves a binary deepfake detection accuracy of 96.3% (outperforming SOTA by over 14.1%) and an attribution mAP of 91.0% (surpassing competitors by over 29.3%) on GenVidBench. Additionally, PDMF demonstrates robust generalization under few-shot conditions.

1 INTRODUCTION

The rapid evolution of video generative technologies (Studio, 2024; Xing et al., 2024; Chen et al., 2023; Ma et al., 2025a; Guo et al., 2024; Zheng et al., 2024; HaCohen et al., 2024; Kong et al., 2024; Wan et al., 2025) creates unprecedentedly numerous AI-generated videos with high quality and realism, easily making imagination become reality. However, as the visual differences between real and AI-generated videos gradually diminish, AI-generated videos are easily misused for exaggerating the truth, prompting false information, manipulating identity information (Karamchand, 2025; Pang et al., 2024b; Spotlight, 2025), *etc.*, thereby posing considerable threats to cybersecurity and social governance.

Much work has focused on deepfake video detection by exploiting AI-generated artifacts by exploring spatial features (Li & Lyu, 2018; Rossler et al., 2019; Frank et al., 2020) and temporal features (Sabir et al., 2019; Ma et al., 2025b; Ji et al., 2024; Xue et al., 2025) from video frames. Despite excellent performance in deepfake video detection, it is insufficient to merely confirm whether the video is real or fake, and tracing the source generators of AI-generated videos has become more paramount from the perspective of criminal tracing and legal supervision.

To this end, research of an effective video source attribution method has received attention. The active attribution methods (Hu et al., 2025a;b; Zhang et al., 2019; Luo et al., 2023; Zhang et al., 2023; Fernandez et al., 2024) embed the model's ownership information into the video, and show significant limitations in complex operations and affecting visual quality. While passive attribution is more effective, most methods (Pang et al., 2024a; Vahdati et al., 2024; Kundu et al., 2025a) are based on a supervised classification paradigm. Despite excellent performance on seen generators during training, these methods entails high costs, massive data and limited generalizability in the few-shot scenario. Besides, few works explore video source attribution from the multimodal perspective and most mutimodal-based methods explore the mismatch between lip-movements and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

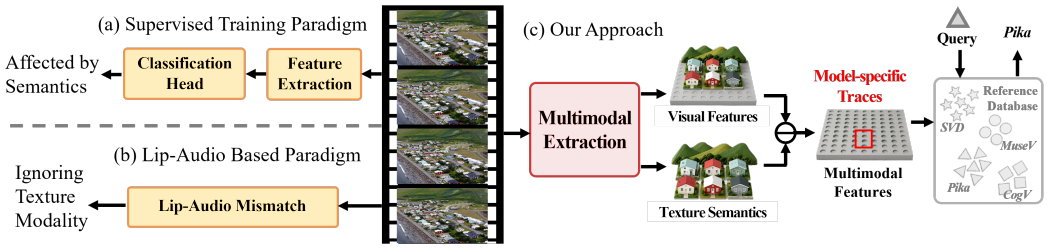


Figure 1: **Motivation.** (a) Existing supervised training paradigm is easily affected by semantics. (b) Existing lip-audio based paradigm ignores the texture modality. (c) Our approach extracts model-specific traces through multimodal features, which accurate, lightweight and scalable.

audio. Therefore, it is of great significance to propose a multimodal approach for video attribution and a paradigm that endows strong capabilities of generalization in the few-shot scenario.

Motivated by these, we propose a prompt-decoupled multimodal framework accompanied with a retrieval paradigm called PDMF that attains much effectiveness on unseen generators, as shown in Fig. 1. We dive into current video generative models, and find that although current generators achieve semantic high fidelity, they generally overlook the rationality of pixel-level statistical signals, leaving behind artifacts. By exploiting the artifacts left by the generator through features extraction, it attains much progress in AI-generated image detection and attribution, but these artifacts are easily affected by semantics, causing the method to overfit to visual information rather than truly distinguishing the generator. In order to eliminate semantic interference and expose underlying model-specific features, we use advanced Vision-Language Models to extract two different granularity feature streams from video instances. Subsequently, we filter out the shared semantic context through a cross modal-residual subtraction mechanism, leaving behind residual noise that implicitly contains generator-specific artifacts.

In addition, we introduce a highly scalable database registration paradigm by transforming the AI-generated video attribution into instance retrieve tasks. The whole system relies merely on a tiny database, in which each generators only needs to store a few model-specific samples, and we only need to attribute the query video to the most similar samples in referencing database. It is crucial that our framework does not require time-consuming retraining or large-scale data. A very small number of samples of new generators are required for system adaption, making it highly scalable. Extensive experiments on GenVidBench demonstrate that PDMF is capable of AI-generated video detection and source attribution in few-shot scenarios by using minimal time and data.

Contributions: 1) Novel perspective for AI-generated video attribution: We innovatively transform AI-generated video attribution into instance retrieve tasks, improving the capability of retrieving unseen models in the few-shot scenario. **2) Multimodal solution to expose generator-specific artifacts:** We effectively expose generator-specific artifacts by fusing multimodal features and remove semantics. **3) Superior performance on few-shot AI-generated video attribution:** Our approach outperforms the state-of-the-art approach on few-shot AI-generated video attribution by 23.7% Rank-1 accuracy and 29.3% mAP by using only 100 videos per class and 2-10 times less time than alternative methods.

2 Related Work

2.1 AI-Generated Video Detection

AI-generated video detection aims to identify whether a video is real or AI-generated. Early research focuses on extracting static spatial features, including image processing (Li & Lyu, 2018), CNN-based (Rossler et al., 2019), and frequency-domain (Frank et al., 2020) methods, but they neglect the temporal dynamics of videos. To overcome the lack of temporal information, some works tend to process the spatial and temporal information in order (Sabir et al., 2019). They first extract the forgery traces at the frame level, and subsequently capture the inconsistency between frames (Vahdati et al., 2024; Ma et al., 2025b). Other advancements utilize state space models (Chen

et al., 2024b), uncertainty-aware representation learning (Kundu et al., 2025b), and transformer-based framework (Yu et al., 2021) to model the temporal evolution. The other works often employ a parallel structure (*e.g.*, AIGVDet (Bai et al., 2024) employs two sub-detectors, DuB3D (Ji et al., 2024) utilizes a dual-branch 3-D Transformer) to handle spatial artifacts and temporal information simultaneously. Xue et al. (2025) utilizes high-order optical flow residuals, in order to amplify subtle anomalies hidden in frames.

2.2 AI-Generated Image/Video Attribution

With the rapid advancement of video generation techniques, AI-generated visual attribution have emerged as a crucial topic in the fields of cybersecurity and judicial forensics. The research of attribution mainly concentrates on images. RONAN (Wang et al., 2023c) utilizes reverse-engineering and reconstruction loss to trace the source generators. In order to overcome limitations of merely recognizing seen generators, Girish *et al.* (Girish et al., 2021) and Yang *et al.* (Yang et al., 2023) expand the closed-set methods to open-set scenarios. Bongini (Bongini et al., 2025) further propose a training-free deepfake image attribution methods by comparing the distance between original and reconstructed images in the CLIP feature space. However, image attribution methods can hardly extend to video attribution (Vahdati et al., 2024), for neglecting temporal dynamics and dependencies between video frames. The exploration of generative model attribution in the field of video is quite limited. VGMShield (Pang et al., 2024a) captures spatial-temporal dynamic inconsistencies between video frames for attribution by utilizing a masked autoencoder as the backbone network. Additionally, several works pioneer the issue of scared annotated data (Kundu et al., 2025a) and the training-free paradigm (Wang et al., 2026). However, most existing approaches of video attribution are confined to a single modality, our approach fully integrates multimodal representations.

2.3 AI-generated Multimodal Video Detection

Most multi-modal methods for AI-generated video detection compare the differences between audio and vision modalities, especially the mismatch between lip movements and audio. Li et al. (2026) combine single-modal independent detection with audio-vision multimodal fractional fusion, achieving impressive performance on deepfake video detection and localization. CAD (Du et al., 2025) utilizes audio-vision alignment to detect high-level semantic dissonance. Anshul et al. (2025) further extend the single audio-vision cross-modal alignment to a two-stage multimodal framework, demonstrating strong performance on generalization temporal localization. However, these works only use audio and visual modalities, ignoring the contribution of other modalities (such as text). Besides, multimodal technology is rarely applied in video source attribution. We attempt to integrate text and vision modalities for video source attribution.

3 Motivation and Problem Formulation

Most existing methods for AI-generated video detection and attribution rely on single modal visual analysis (He et al., 2024; Pang et al., 2024a; Kundu et al., 2025a), which exposes artifacts left by the generator through visual feature processing, with subsequent supervised training. These methods are highly susceptible to semantic interference, as the model is prone to overfitting to visual content, resulting in insufficient generalization ability in open-world or few-shot scenarios. Conversely, existing multimodal video detection methods (Katamneni & Rattani, 2024; Li et al., 2026; Du et al., 2025; Anshul et al., 2025) primarily align the features of different modalities to discover inconsistencies, suffering from poor interpretability. Further, most of these methods focus on visual and audio modalities. These multimodal methods ignore the crucial textual modality and require significant computational overhead.

Recent studies have shown that although current generative models achieve high fidelity at the semantic level, generator-specific traces inevitably persist within underlying pixel-level statistical signals. In order to eliminate semantic interference and accurately capture these traces, it is necessary to introduce a multimodal method that fuses visual and textual modalities. We project textual features into semantic information and use the cross-modal residual subtraction mechanism to explicitly filter out semantic context shared with visual features, thereby fully exposing generator-specific traces

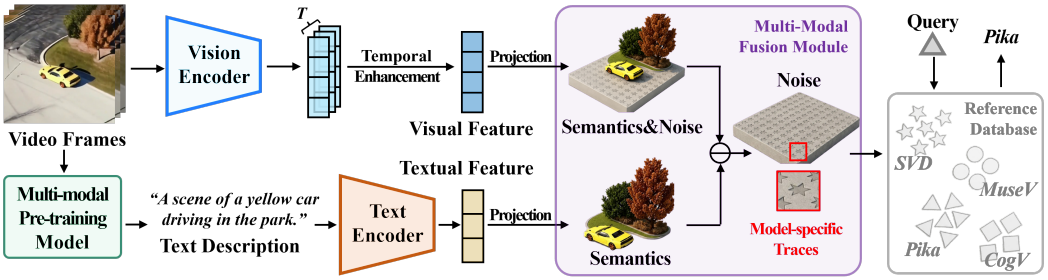


Figure 2: **Overview of our approach.** Firstly, we extract raw visual features from sampled video frames using a visual encoder, and enhance temporal dependencies. We also generate descriptive subtitles as textual aids using a multimodal pre-trained model. Secondly, we use a cross-modal residual subtraction mechanism to filter out shared semantic context, successfully exposing generator-specific artifacts. Finally, we establish a highly scalable database registration paradigm that transforms the video source attribution into an instance retrieval process based on cosine similarity.

hidden in video noise. By further introducing a retrieval database to reduce the required data, it makes AI-generated video detection and attribution efficient, generalizable, and interpretable.

Input: Given a video instance, its input contains two different modalities: (1) Visual video stream: The L frame image sequence obtained by continuously sampling $V = \{f_1, f_2, \ldots, f_L\}$. (2) Textual assistance stream: Textual description used to generate this video T_{raw}. If there is no textual description, the text can be reversed from the video.

Modeling: Assuming there is a set of known video generators in the system $\mathcal{G} = \{G_0, G_1, \ldots, G_j, \ldots, G_M\}$ (G_0 represents $Real$, and $G_1, \ldots, G_M$ denote generative classes), we need to identify generator-specific artifact features F^j by modeling visual and textual inputs:

$$F_j = M(V, T_{\text{raw}}). \tag{1}$$

Furthermore, a reference database $\mathcal{D}$ is established where only a small number of extracted feature samples (*i.e.*, N) are registered for each generator category $G_j \in \mathcal{G}$:

$$\mathcal{D} = \left\{ \{F_i^j\}_{i=1}^{N} \right\}_{j=1}^{M}. \tag{2}$$

This database has low construction cost and excellent scalability.

Output: During inference, for any query video V_{query}, the same pipeline is employed to extract the generator-specific feature F_{query}. Subsequently, calculate the cosine similarity between F_{query} and all stored features in reference database $\mathcal{D}$, and find the feature with the highest similarity F_i^j in the database. (1) Source attribution: According to the class G_j the matched feature F_i^j belongs to, let $V_{query} \in G_j$. (2) Deepfake detection: If $j = 0$, let $V_{query} \in Real$; otherwise, let $V_{query} \in Fake$.

4 METHOD

We propose Prompt-Decoupled MultiModal Forensics (PDMF), which is shown in Fig. 2. PDMF is composed of four components: Visual Extraction and Temporal Enhancement, Textual Assistance, Multimodal Fusion, and Retrieval Perspective for Video Attribution. Details are described below.

4.1 MULTIMODAL FORENSIC FEATURES

For each video instance V, we first perform continuous sampling to obtain a sequence containing L frames to capture temporal motion dynamic information, denoted as $V = \{f_1, f_2, \ldots, f_L\}$.

To extract high-level semantic and spatial representations from these raw frames, we employ a pre-trained model (*e.g.*, CLIP) as the visual extraction backbone network. Each independent video frame f_i is mapped into a fixed dimensional visual vector through the visual encoder, and finally

concatenated along the time dimension to output the raw temporal visual feature $V_{\text{raw}} \in \mathbb{R}^{B\times L\times D}$, where B represents the batch size and D is the visual feature dimension for each frame.

Although visual features have excellent capabilities of video source attribution, text descriptions, especially those derived from videos, can also describe the differences of video source in terms of probability distribution. For instance, some generators tend to render scenes with more adjectives in the text, while others are good at depicting actions with verbs dominating the text.

To avoid several videos not providing corresponding text descriptions, we uniformly use advanced generative visual-language models (*e.g.*, BLIP) to automatically synthesize corresponding text descriptions through video frames. Subsequently, the generated raw text description T_{raw} is tokenized and encoded using the text encoder of a pre-trained model, to output the global textual feature T_G.

4.2 Visual Enhancement with Texture Assistance

Temporal Visual Enhancement: In order to effectively aggregate sequence-level temporal dependencies from the original frame-level embedding V_{raw}, we encapsulates a flexible Temporal-Modeling module to obtain the global visual feature V_G, which employs a standard Transformer encoder layer. Given the raw sequential visual features $V_{raw} \in \mathbb{R}^{B\times L\times D}$, the query Q, key K and value V are all the linear projections of V_{raw}. The Multi-Head Attention (MHA) mechanism, subsequent Feed-Forward Network (FFN), and the temporal pooling aggregation are formulated as:

$$V_L = \text{FFN}(\text{MHA}(Q, K, V)) + V_{raw}, V_G = \frac{1}{L}\sum_{i=1}^{L} V_L[:, i, :], \tag{3}$$

where V_L denotes the temporally enhanced sequence features, and L represents the length of frames. Note that layer normalization and residual connections are implicitly integrated into the standard Transformer encoder layer. Finally, we apply the time averaged pooling operation on the temporal dimension to obtain the global visual representation $V_G \in \mathbb{R}^{B\times D}$.

In order to enhance the assistance of textual features T_G to visual features V_G, we mapped them into a unified space, resulting in T_{emb} and V_{emb}. Then we introduce an infoNCE-based bidirectional contrastive loss to align the visual and textual embeddings. We define visual-to-text alignment loss $\mathcal{L}_{\text{v2t}}$ and text-to-visual alignment loss $\mathcal{L}_{\text{t2v}}$:

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{v}_i^\top \mathbf{t}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{v}_i^\top \mathbf{t}_j/\tau)}, \tag{4}$$

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\mathbf{t}_i^\top \mathbf{v}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{t}_i^\top \mathbf{v}_j/\tau)}, \tag{5}$$

where τ is learnable logit temperature scale. $\mathbf{v}_i$ and $\mathbf{t}_i$ represents the L_2 normalized visual and textual embeddings of the i-th sample in a batch of samples of size B, ensuring that matched video-text pairs are tightly pulled together. The contrastive learning loss $\mathcal{L}_{\text{align}}$ is computed as:

$$\mathcal{L}_{\text{align}} = (\mathcal{L}_{\text{v2t}} + \mathcal{L}_{\text{t2v}})/2. \tag{6}$$

Multimodal Fusion: It is crucial to find an effective approach that seamlessly integrates the visual feature V_{emb} and textual feature T_{emb}. To achieve this purpose, a straightforward idea is to directly concatenate two modalities into a joint dual-channel vector space, or adopt a multi-head cross-modal attention mechanism. However, these methods have limited capabilities of completely eliminating semantic information from visual features and retaining generator-specific artifact features.

Considering that the generator-specific artifacts that can distinguish each generator is generally hidden in the noise area of the video frame rather than the semantic area, we project the textual feature T_{emb} onto the visual feature space to obtain the semantic representation of the visual feature, and then subtract this representation from the complete visual feature to obtain the residual feature, which is filled with the model-specific artifacts. The fused feature is formulated as:

$$F_{\text{fuse}} = V_{\text{emb}} - \left(\frac{V_{\text{emb}} \cdot T_{\text{emb}}^\top}{\|T_{\text{emb}}\|_2^2 + \epsilon}\right) T_{\text{emb}}, \tag{7}$$

where ϵ is a small positive constant.

Finally, we project V_{emb}, T_{emb} and F_{fuse} to three classification heads respectively, and three heads are jointly optimized simultaneously. The ultimate objective function jointly supervises the three classification tasks and bidirectional alongside the bidirectional contrastive alignment constraint:

$$\mathcal{L}_{\text{total}} = \text{CE}(\mathbf{l}_{\text{vis}}, Y_{\text{cls}}) + \text{CE}(\mathbf{l}_{\text{tex}}, Y_{\text{cls}}) + \text{CE}(\mathbf{l}_{\text{fuse}}, Y_{\text{cls}}) + \lambda \cdot \mathcal{L}_{\text{align}}, \tag{8}$$

where Y_{cls} is the ground-truth video source label, and λ scales the bidirectional contrastive loss.

4.3 Training and Inference

For practical source attribution, we view attribution as a retrieval task rather than a classification task. This paradigm has strong advantages: On one hand, newly emerged generators can be easily added by registering new samples without retraining the model, making it **highly scalable**. On the other hand, it supports evidence-based attribution, and the retrieved reference videos provide intuitive support for the prediction results, thereby improving **interpretability and reliability**.

We regard generator-specific fused feature F_{fuse} extracted from video frames as the key basis, so we extract the visual and textual feature of each video to obtain the generator-aware fused feature. Specifically, we maintain a reference database of AI-generated videos, by using the fused feature F_{fuse}, denoted as $\mathcal{D} = \{F_1^j, \ldots, F_i^j, \ldots, F_N^M\}$, where j denotes the index of generator G_j (M generators in total) and i denotes the index of a video sample (N video samples per class). For a query video, we apply the same processing and perform attribution via retrieval in reference database.

5 Experiments

5.1 Experimental Setup

Dataset: GenVidBench (Ni et al., 2025) is a recently introduced and massive benchmark for AI-generated video detection, which encompasses over 100,000 video-text pairs and original generation prompts. It bridges real-world videos (sourced from HD-VG (Wang et al., 2023b) and Vript (University, 2024)) and AI-generated variations generated by 8 diverse mainstream AI video generators (T2VZ (Khachatryan et al., 2023), MS (Wang et al., 2023a), VC2 (Chen et al., 2024a), Pika (pika, 2024), SVD (Blattmann et al., 2023), MuseV (Xia et al., 2024), Mora (Yuan et al., 2024), and CogVideo (Hong et al., 2022)). Given its diverse coverage across real and AI-generated videos, it is well suited for AI-generated video attribution. In this task, we treat all samples from HD-VG and Vript as a singular *Real* category during attribution.

Evaluation Metrics: To evaluate AI-generated video attribution, we employ Rank-1 accuracy and mean Average Precision (mAP). Rank-1 accuracy represents the percentage of query videos of which the highest-probability prediction correctly matches the ground-truth source. The mAP metric calculates the mean of the average precision across all queries. For deepfake detection, we use accuracy as the evaluation metric.

Implementation Details: We sample 32 consecutive frames from each video, and temporal looping is applied to videos with fewer than 32 frames. Specifically, a pre-trained CLIP model is used to extract the original visual features and textual features, and the time series of visual features is enhanced through an adaptive time modeling layer. Independent linear classification heads are deployed to simultaneously supervise the flow of visual, textual, and fused features. The framework undergo end-to-end optimization within 30 training epochs, with a bidirectional contrastive learning loss weight of $\lambda = 0.1$. After training, the network weights are frozen and a selected subset of videos is used to extract features to fill the reference database. During retrieval, query videos undergo the same process, and cosine similarity is computed against all stored features for source attribution.

5.2 Experimental Results

We utilize constructed baselines, deepfake image and video detection methods for comparison: (1) **Constructed baselines**: We extract frame-level features using ResNet-50 (He et al., 2016) and aggregate them via temporal average pooling. Besides, R3D-18 (Tran et al., 2018) is applied to continuous frames to capture spatiotemporal representations. (2) **Deepfake image detection methods**:

Shot	Method	Real		T2VZ		MS		VC2		Pika		SVD		MuseV		Mora		CogV		Avg.	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
5-shot	ResNet-50 (He et al., 2016)	6.5	19.7	10.1	19.7	26.1	30.7	15.6	26.7	24.6	37.1	20.6	32.6	27.6	36.7	16.6	24.2	28.6	34.2	18.3	27.9
	R3D-18 (Tran et al., 2018)	30.2	42.8	31.2	37.4	54.3	50.6	18.6	32.4	28.1	37.1	12.1	22.5	25.6	35.0	29.2	34.1	48.7	44.9	30.9	37.4
	ESSP (Chen et al., 2024c)	22.1	30.3	12.1	24.8	35.2	48.7	20.6	30.4	30.2	30.8	20.1	29.1	10.6	23.7	19.6	32.6	8.0	18.8	19.8	29.9
	LOTA (Wang et al., 2025)	32.7	38.5	28.1	30.3	77.4	70.3	48.7	42.3	57.3	51.6	38.7	40.7	31.7	38.0	18.1	30.4	68.8	55.5	44.6	44.2
	PiD (Fu et al., 2025)	27.1	36.0	47.2	41.7	32.7	39.1	11.6	23.3	26.6	33.8	35.2	39.6	35.2	34.7	32.7	39.4	68.3	54.9	35.2	38.1
	DeMamba (Chen et al., 2024b)	20.1	-	31.7	-	9.6	-	15.1	-	18.1	-	27.1	-	23.6	-	28.6	-	44.7	-	24.3	-
	UNITE (Kundu et al., 2025b)	7.8	-	66.8	-	64.8	-	30.2	-	10.6	-	2.5	-	78.9	-	50.8	-	34.2	-	35.1	-
	VGMShield (Pang et al., 2024a)	31.6	-	26.5	-	67.4	-	15.3	-	67.4	-	13.3	-	26.5	-	1.0	-	83.7	-	36.9	-
	PDMF	64.6	70.2	95.5	94.5	31.0	43.5	16.5	28.9	27.6	42.4	28.2	39.3	50.9	59.2	8.2	20.0	77.4	77.2	44.4	52.8
10-shot	ResNet-50 (He et al., 2016)	33.4	28.9	13.6	14.8	26.1	20.7	15.6	18.6	29.2	23.0	15.6	17.3	16.1	18.7	17.1	17.6	41.7	31.5	24.2	32.0
	R3D-18 (Tran et al., 2018)	17.1	25.5	45.2	45.8	28.1	34.6	30.2	37.3	43.7	44.1	21.6	31.5	65.3	63.1	12.6	22.5	84.4	76.6	38.7	42.3
	ESSP (Chen et al., 2024c)	18.6	16.2	16.1	16.8	59.8	50.0	31.2	25.3	27.6	20.5	10.1	12.8	19.1	19.9	12.6	18.0	37.2	30.4	25.8	23.3
	LOTA (Wang et al., 2025)	34.2	38.0	52.8	48.1	78.4	74.9	51.3	51.5	41.2	42.2	38.2	35.9	40.2	44.4	29.6	32.0	81.4	63.5	49.7	47.8
	PiD (Fu et al., 2025)	34.2	37.4	43.7	44.5	38.7	39.1	26.6	28.6	50.3	44.8	36.2	39.4	40.7	41.8	38.7	38.3	67.8	52.6	41.9	40.7
	DeMamba (Chen et al., 2024b)	12.6	-	60.8	-	22.6	-	24.6	-	27.1	-	22.6	-	14.1	-	11.6	-	29.6	-	25.1	-
	UNITE (Kundu et al., 2025b)	47.0	-	48.2	-	26.1	-	27.1	-	18.6	-	14.1	-	72.9	-	37.2	-	76.5	-	40.8	-
	VGMShield (Pang et al., 2024a)	38.8	-	68.4	-	21.4	-	10.2	-	63.3	-	17.4	-	7.1	-	48.0	-	63.3	-	37.5	-
	PDMF	57.3	71.9	100.0	99.9	53.8	61.4	28.6	40.0	51.8	60.4	48.5	52.0	60.1	53.7	44.5	55.2	58.5	65.0	55.9	62.2
100-shot	ResNet-50 (He et al., 2016)	34.4	32.9	22.6	21.5	39.2	31.4	27.6	24.2	27.6	24.7	23.1	19.6	12.6	14.8	22.1	20.6	45.7	34.8	28.9	25.8
	R3D-18 (Tran et al., 2018)	43.0	45.4	73.4	66.1	47.7	48.9	42.7	42.9	53.8	53.4	21.1	27.2	48.7	49.6	36.2	38.4	92.0	87.8	50.9	51.1
	ESSP (Chen et al., 2024c)	21.6	20.2	14.6	15.6	55.3	49.4	28.6	26.1	14.1	18.1	21.6	20.4	18.1	18.2	23.1	22.6	38.7	34.6	26.2	25.0
	LOTA (Wang et al., 2025)	59.3	54.9	79.0	70.3	96.0	92.9	64.3	58.9	55.8	57.6	46.2	46.5	50.3	51.3	40.7	41.9	85.9	81.2	64.1	61.7
	PiD (Fu et al., 2025)	44.2	45.2	70.4	63.5	58.3	54.7	39.7	42.8	55.8	54.6	54.3	51.3	51.3	52.8	42.2	43.9	88.9	81.1	56.1	54.4
	DeMamba (Chen et al., 2024b)	17.6	-	61.3	-	35.7	-	36.7	-	40.7	-	21.1	-	28.6	-	23.1	-	37.7	-	33.6	-
	UNITE (Kundu et al., 2025b)	5.3	-	88.9	-	79.4	-	15.6	-	66.3	-	8.0	-	56.3	-	85.4	-	85.4	-	49.6	-
	VGMShield (Pang et al., 2024a)	38.8	-	98.0	-	83.7	-	74.5	-	69.4	-	41.8	-	84.7	-	49.0	-	98.0	-	70.9	-
	PDMF	89.5	93.1	100.0	100.0	85.8	90.3	76.8	82.1	76.4	82.9	89.6	91.5	89.0	92.2	96.3	97.5	87.1	89.7	87.8	91.0

Table 1: **Source attribution evaluation.** We compare the attribution performance of our approach against other competing methods under 5-shot, 10-shot and 100-shot, respectively. The best results are highlighted in bold, and the second-best results are underlined.

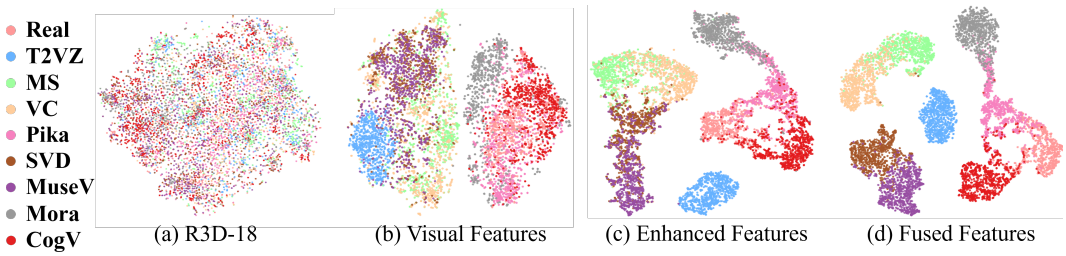
Method	HD-VG	Vript	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
ResNet-50 (He et al., 2016)	19.0	35.4	88.9	84.8	87.7	86.6	83.0	82.4	87.9	90.9	80.0
R3D-18 (Tran et al., 2018)	32.2	42.2	96.5	92.0	88.9	86.4	68.3	79.4	77.8	100.0	76.4
ESSP (Chen et al., 2024c)	26.1	21.6	85.9	92.3	86.1	83.4	85.9	82.4	90.4	81.4	73.6
LOTA (Wang et al., 2025)	57.8	53.8	99.5	100.0	99.4	87.4	94.0	91.9	89.1	100.0	87.3
PiD (Fu et al., 2025)	50.7	65.3	100.0	100.0	100.0	93.0	99.1	98.4	95.3	99.8	90.2
DeMamba (Chen et al., 2024b)	54.8	59.3	94.0	96.1	95.4	96.5	62.7	41.2	82.9	90.9	76.4
UNITE (Kundu et al., 2025b)	13.6	12.6	100.0	100.0	100.0	100.0	98.5	99.0	99.4	98.6	82.2
PDMF (10-shot)	74.2	83.7	100.0	100.0	100.0	55.0	100.0	100.0	78.7	67.2	85.9
PDMF (100-shot)	85.9	88.6	100.0	100.0	100.0	92.8	100.0	100.0	99.6	95.6	96.3

Table 2: **Deepfake detection evaluation.** We compare the deepfake video detection accuracy with competing methods under 100-shot scenarios.

The raw video frames are enhanced using mainstream deepfake detection methods from ESSP (Chen et al., 2024c), LOTA (Wang et al., 2025), or PiD (Fu et al., 2025). Subsequently, these processed frames are fed into R3D-18 to extract features for constrcuting the database. (3) **Deepfake video detection methods**: We adapt the DeMamba (Chen et al., 2024b) and UNITE (Kundu et al., 2025b) by substituting the binary classification heads with a nine-category classifier. (4) **Video Attribution methods**: We apply the video attribution method VGMShield (Pang et al., 2024a) for comparison.

Source Attribution Evaluation: We evaluate PDMF on GenVidBench under three settings: 5-shot, 10-shot, and 100-shot, by utilizing 5, 10, and 100 videos per class for training, respectively. In Tab. 1, PDMF outperforms all competing methods. In the 100-shot setting, PDMF delivers a Rank-1 accuracy of over 87% and an mAP surpassing 91%, surpassing competitors by over 16.9% Rank-1 and 29.3% mAP. More specifically, PDMF achieves balanced and excellent performance across all categories, while alternative methods tend to perform poorly on certain categories, which indicates their limited generalization capabilities on video attribution. Even in the 5-shot setting, other methods have few capabilities of source attribution with even less than 30% mAP, while PDMF still significantly outperforms other methods by achieving 52.8% mAP.

Deepfake Detection Evaluation: PDMF is also capable of binary deepfake detection, by using a few samples for training. We train all baselines and competitive methods using 100 videos per class for fair comparison, and the deepfake detection results are shown in Tab. 2. PDMF comprehensively outperforms all other competitive methods by achieving an impressive average accuracy of 96.3%, surpassing UNITE by 14.1% and DeMamba by 19.9%. Remarkably, even merely 10 videos per class for training, PDMF consistently outperforms other methods which use 100 videos per class, yielding an average accuracy of 85.9%. Moreover, other methods exhibit significantly lower accuracy on the real subsets (HD-VG and Vript), hardly striking a balance between real and AI-generated videos. In contrast, PDMF maintains exceptionally balanced performance across all datasets and categories, showing the superior capability of binary deepfake video detection.



(a) R3D-18 (b) Visual Features (c) Enhanced Features (d) Fused Features

Figure 3: **Visualization of video features distribution from nine classes in GenVidBench.** We provide the video features distribution generated by four different methods.

Vision	Text	Fusion	Real	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
✓			81.1	100.0	87.1	78.0	62.4	78.8	83.8	73.6	89.5	81.6
	✓		21.4	21.6	15.6	15.9	15.6	17.2	16.1	15.6	16.4	17.3
✓	✓		94.4	100.0	90.2	82.1	81.7	76.8	76.0	98.0	82.8	86.9
✓	✓	✓	93.1	100.0	90.3	82.1	82.9	91.5	92.2	97.5	89.7	91.0

Table 3: **Ablation studies.** We remove three branches of our approach to verify the effectiveness of each module: visual branch, textual branch, and fusion branch.

Visualizations of extracted features: We exhibit the visualizations of video features extracted from 9 classes in GenVidBench by using four different methods. As shown in Fig. 3, features extracted through pre-trained R3D-18 (a) are mixed together, and have few capabilities of distinguishing each classes. By using pre-trained vision encoder of CLIP (b), video features begin to show characteristics of clustering by class, but there are still several features mixed with each other, which is a manifestation of underfitting. Further, applying enhancement to (b), the enhanced visual features (c) exhibit clearer differentiation between classes, but there is a tendency for overfitting. PDMF combines visual and textual features (d), and the non-semantic generator-specific features perfectly separate each class from another, with strong capabilities of source attribution.

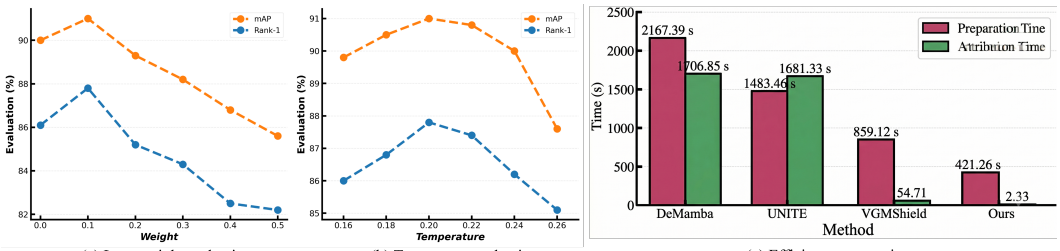
5.3 Ablations and Analysis

Ablation Studies: To validate the effectiveness of each module of our approach, four experimental setups are considered: (1) using the visual branch only; (2) using the textual branch only; (3) using both visual and textual branches; and (4) using all branches, as in our approach. The results are shown in Tab. 3. Using textual branch only yields the poorest attribution mAP of 17.3%, indicating that textual features alone do not contain the ability to distinguish between different categories of videos. Using visual branch only achieves reasonable attribution mAP of 81.6%, demonstrating the strong capability of video attribution in visual features. When both visual and textual features are applied, the attribution mAP improves to 86.9%, which indicates that multimodal contrastive learning can effectively narrow the distance between visual and textual spaces, achieving preliminary complementarity of features. Finally, when the fusion branch is also applied, the attribution mAP reaches the peak of 91.0%, which proves that introducing a deep multimodal fusion module can further capture complex cross-modal interaction features and mine model-specific traces for video attribution.

Analysis of Fusion Methods: In Sec. 4.2, we apply residual mechanism to fuse the visual and textual features. We study the impacts of different fusion methods in Tab. 4, which includes: (1) Concat: We directly concatenate two modalities into a joint dual-channel vector space; (2) Attention: We Adopt a multi-head cross-modal attention mechanism, using textual embeddings T_{emb} as the query Q, visual embeddings V_{emb} as the key K and value V; (3) Residual: Our approach. Results show that Residual achieves the best overall performance of 91.0% mAP, slightly higher than Attention and Concat. Obviously, Concat only mechanically cascades the two features, lacking deep modal interaction. Despite the cross-modal relationships built by cross-attention, Attention introduces redundancy and noise during fusion. In contrast, Residual strips common semantics while maintaining visual features that contain model-specific artifacts, resulting in better performance.

Fusion	Real	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
Concat	90.4	100.0	85.2	89.0	83.9	88.2	83.4	98.2	90.7	89.9
Attention	87.6	100.0	91.8	75.0	77.2	99.6	99.9	96.8	84.9	90.3
Residual	93.1	100.0	90.3	82.1	82.9	91.5	92.2	97.5	89.7	91.0

Table 4: **Analysis of fusion methods.** We compare three fusion methods, Concat, Attention, and Residual, to find the most suitable one.



(a) Loss weights selection (b) Temperature selection (c) Efficiency comparison

Figure 4: **Analysis of hyper-parameters and the efficiency of video attribution.** We demonstrate the relationship between Rank-1 and mAP on (a) loss weights and (b) temperature in contrastive learning. Besides, We compare the (c) time consumption of PDMF with other video detection methods during preparation and attribution.

Analysis of Loss Weights: To study the impact of using different proportions of cross-entropy loss and contrastive learning loss, we apply different weights for contrastive learning loss from 0 to 0.5. As shown in Fig. 4 (a), the mAP attains the peak of 91.0% when the weight of contrastive learning loss is set to 0.1. This is attributed to the trade-off in the contrastive loss weight: excessive contrastive loss leads to a deviation from the original classification focus, while reduction of contrastive loss fails to ensure sufficient cross-modal alignment.

Analysis of Temperature Coefficient: The temperature coefficient in contrastive learning has a significant impact on distancing samples of different categories and enhancing the capability of attribution. As shown in Fig. 4 (b), we adjust the temperature parameter from 0.16 to 0.26, and the attribution mAP first increases and then decreases with temperature changes, reaching the best performance of 91.0% at 0.20. This is mainly attributed to the temperature controlling the sensitivity of the contrastive loss function to distinguish negative samples: an excessively low temperature overly focuses the model on a tiny fraction of hard negative samples, resulting in overfitting. Conversely, a higher temperature oversmooths the probability distribution across the similarity matrix, losing the model's sensitivity to inter-class discrimination of weak generated traces. Therefore, a moderate temperature perfectly balances the two aspects, thereby achieving strong attribution performance.

Analysis of the efficiency of video attribution: We compare the attribution efficiency of PDMF and other mainstream AI-generated video detection and attribution methods (DeMamba, UNITE, and VGMShield). For each class, 100 videos are utilized for testing in the 100-shot scenario. We analyze the time consumption in the stage of preparation (contains feature extraction, training and database construction) and attribution. Fig. 4 (c) shows that PDMF only takes 421.26s in the preparation stage and 2.33s in the attribution stage, which is 2-10 times faster than other methods.

6 Conclusion

We propose a novel multimodal retrieval framework called PDMF addressing the issues of AI-generated video detection and attribution. By utilizing visual-language models to extract features and employing cross-modal residual subtraction, PDMF explicitly strips away shared semantic context, effectively exposing subtle, generator-specific artifacts hidden in video noise. Extensive evaluation of the GenVidBench dataset shows that PDMF achieves excellent performance, with a binary deepfake detection accuracy of 96.3% and an attribution mAP of 91.0% by using merely 100 samples for training and 2-10 times less time, significantly outperforming state-of-the-art baselines. It is crucial that PDMF provides competitive generalization in demanding few-shot scenarios.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and assistance with code development. All AI-assisted content were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not introduce new training data or collect human-subject data. The proposed method inherits potential risks and biases from the underlying pretrained generative models.

REFERENCES

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13826–13836, 2025.

Jianfa Bai, Man Lin, and Gang Cao. Ai-generated video detection via spatio-temporal anomaly learning, 2024. URL https://arxiv.org/abs/2403.16638.

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.

Pietro Bongini, Valentina Molinari, Andrea Costanzo, Benedetta Tondi, and Mauro Barni. Training-free source attribution of ai-generated images via resynthesis, 2025. URL https://arxiv.org/abs/2510.24278.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024a.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024b.

Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024c.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023.

Yuxuan Du, Zhendong Wang, Yuhao Luo, Caiyong Piao, Zhiyuan Yan, Hao Li, and Li Yuan. Cad: A general multimodal framework for video deepfake detection via cross-modal alignment and distillation, 2025. URL https://arxiv.org/abs/2505.15233.

Pierre Fernandez, Hady Elsahar, I Zeki Yalniz, and Alexandre Mourachko. Video Seal: Open and Efficient Video Watermarking. *arXiv preprint arXiv:2412.09492*, 2024.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pp. 3247–3258. PMLR, 2020.

Xinghe Fu, Zhiyuan Yan, Zheng Yang, Taiping Yao, Yandan Zhao, Shouhong Ding, and Xi Li. Pid: Generalized ai-generated images detection with pixelwise decomposition residuals. In *International Conference on Machine Learning*, 2025.

Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14094–14103, 2021.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations*, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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*, pp. 770–778, 2016.

Peisong He, Leyao Zhu, Jiaxing Li, Shiqi Wang, and Haoliang Li. Exposing ai-generated videos: A benchmark dataset and a local-and-global temporal defect based detection method, 2024. URL https://arxiv.org/abs/2405.04133.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking. In *International Conference on Learning Representations*, 2025a.

Xuming Hu, Hanqian Li, Jungang Li, Yu Huang, Shuliang Liu, Qi Zheng, Junhao Chen, and Aiwei Liu. VideoMark: A Distortion-Free Robust Watermarking Framework for Video Diffusion Models. *arXiv preprint arXiv:2504.16359*, 2025b.

Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. Distinguish any fake videos: Unleashing the power of large-scale data and motion features, 2024. URL https://arxiv.org/abs/2405.15343.

Gopalakrishna Karamchand. Detecting the Abuse of Generative AI in Cybersecurity Contexts: Challenges, Frameworks, and Solutions. *Journal of Data Analysis and Critical Management*, 1(03): 1–12, 2025.

Vinaya Sree Katamneni and Ajita Rattani. Contextual cross-modal attention for audio-visual deepfake detection and localization, 2024. URL https://arxiv.org/abs/2408.01532.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*, 2024.

Rohit Kundu, Vishal Mohanty, Hao Xiong, Shan Jia, Athula Balachandran, and Amit K. Roy-Chowdhury. Saga: Source attribution of generative ai videos, 2025a. URL https://arxiv.org/abs/2511.12834.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K. Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content, 2025b. URL https://arxiv.org/abs/2412.12278.

Qingcao Li, Miao He, Liang Yi, Qing Wen, Yitao Zhang, Hongshuo Jin, Peng Cheng, Zhongjie Ba, Li Lu, and Kui Ren. Divide and conquer: Multimodal video deepfake detection via cross-modal fusion and localization, 2026. URL https://arxiv.org/abs/2602.00209.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

Xiyang Luo, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. DVMark: A Deep Multiscale Framework for Video Watermarking. *IEEE Transactions on Image Processing*, 2023.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-Video-T2V Technical Report: The Practice, Challenges, and Future of Video Foundation Model. *arXiv preprint arXiv:2502.10248*, 2025a.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency, 2025b. URL https://arxiv.org/abs/2402.02085.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video, 2025. URL https://arxiv.org/abs/2501.11340.

Yan Pang, Baicheng Chen, Yang Zhang, and Tianhao Wang. VGMShield: Mitigating Misuse of Video Generative Models. *arXiv preprint arXiv:2402.13126*, 2024a.

Yan Pang, Aiping Xiong, Yang Zhang, and Tianhao Wang. Towards Understanding Unsafe Video Generation. *arXiv preprint arXiv:2407.12581*, 2024b.

pika. Pika. Online, 2024. Available:https://pika.art/home.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3 (1):80–87, 2019.

Spotlight. AI-generated footage falsely used to show 'aftermath' of Philippines earthquake. *https://spotlight.ebu.ch/*, 2025.

Morph Studio. Morph studio. https://www.morphstudio.com/, 2024.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Shanghai Jiao Tong University. Vript: A video is worth thousands of words. Online, 2024. Available:https://github.com/mutonix/Vript.

Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4397–4408, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Chao Wang, Zijin Yang, Yaofei Wang, Yuang Qi, Weiming Zhang, Nenghai Yu, and Kejiang Chen. Swift: Sliding window reconstruction for few-shot training-free generated video attribution, 2026. URL https://arxiv.org/abs/2603.08536.

Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. Lota: Bit-planes guided ai-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17246–17255, October 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023b.

Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems*, 36: 74478–74500, 2023c.

Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. Musev: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising. *arxiv*, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Xi Xue, Kunio Suzuki, Nabarun Goswami, and Takuya Shintate. Video forgery detection with optical flow residuals and spatial-temporal consistency, 2025. URL https://arxiv.org/abs/2508.00397.

Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15856–15865, 2023.

Bingyao Yu, Wanhua Li, Xiu Li, Jiwen Lu, and Jie Zhou. Frequency-aware spatiotemporal transformers for video inpainting detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8188–8197, 2021.

Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework, 2024.

Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust Invisible Video Watermarking with Attention. *arXiv preprint arXiv:1909.01285*, 2019.

Yulin Zhang, Jiangqun Ni, Wenkang Su, and Xin Liao. A Novel Deep Video Watermarking Framework with Enhanced Robustness to H. 264/AVC Compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8095–8104, 2023.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All. *arXiv preprint arXiv:2412.20404*, 2024.

SUPPLEMENTARY MATERIAL

A ANALYSIS OF VISUAL ENHANCEMENT

In Sec. 4.2, we apply temporal enhancement via Transformer encoder, so we also evaluate three different visual enhancement methods to improve the representation of visual features: directly temporal pooling, Temporal Convolutional Network (TCN) and our Transformer encoder. Tab. 5 shows the results of applying these three enhancements, where Transformer exhibits overwhelming advantages, with an average mAP of 91.0%, significantly surpassing TCN's 61.8% and Pooling's 60.1%. This performance gap stems from the fact that Pooling merely performs parameter-free averaging, completely losing the temporal evolution clues between frames. Although TCN introduces local correlations and state scanning, its capability of capturing long-range dynamic features with large spans in videos is limited. In contrast, Transformer, with its inherent Self-Attention mechanism, can effectively model global non-linear dependencies across the entire temporal domain, thereby extracting more robust multimodal attribution representations.

Enhancement	Real	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
Pooling	51.1	99.9	58.3	70.7	71.1	35.2	38.2	49.8	66.6	60.1
TCN	70.2	99.2	58.8	59.6	56.6	56.2	47.7	49.5	58.0	61.8
Transformer	93.1	100.0	90.3	82.1	82.9	91.5	92.2	97.5	89.7	91.0

Table 5: **Analysis of visual enhancement.** We compare three visual enhancement methods: Pooling, TCN, and Transformer.

B ANALYSIS OF BRANCH FUSION

We further investigate different methods of fusing three branches in Tab. 6. (1) Predefined weights: We give predefined weights to visual-dominant, text-dominant and modal-fusion branches, maintaining the weights of the first two branches consistent. The best result using predefined weights is only 77.5% mAP, indicating it difficult to find a suitable set of weights to fuse three branches. (2) Adaptive weights: The model adapt the weights for three branches during training, by assign more weights to the branch with stronger gradients during gradient descent. Even the optimal weight combination yields the poor result of 78.4% mAP, suggesting that joint optimization of three branches is infeasible. (3) Independent optimization: We update the weight of each branch independently, and take the average of three branches as the final decision. The superior performance of 91.2% stems from its capability of eliminating gradient interference and negative transfer among three branches. By decoupling optimization, each branch can fully exploit its specific modal features and converge to its optimal space respectively.

Branch Fusion	Real	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
$[0.2, 0.2, 0.6]$	83.0	99.5	62.8	70.4	73.6	53.4	56.7	93.4	81.0	74.9
$[0.3, 0.3, 0.4]$	81.0	100.0	67.8	70.4	78.0	54.4	59.0	96.0	90.9	77.5
$[0.4, 0.4, 0.2]$	74.0	99.8	58.1	57.3	69.8	40.6	60.6	95.1	74.5	70.0
Adaptive Weights	73.4	98.3	71.2	65.4	71.8	84.9	82.5	69.3	89.1	78.4
Independent	93.1	100.0	90.3	82.1	82.9	91.5	92.2	97.5	89.7	91.0

Table 6: **Analysis of branch fusion.** We compare the effects of pre-trained weights, adaptive weights, and independent optimization for branch fusion.

C SENSITIVITY TO REGISTERED SAMPLES PER CLASS

To study PDMF's sensitivity to the number of registered samples per class, we change registered samples from 1 to 50 per class (Tab. 7). As the number of registered samples per class increases, the attribution Rank-1 accuracy slowly rises first and then remains almost unchanged, overall remaining

stable between 85% and 88%. This indicates that PDMF has reasonable sensitivity to the number of registered samples per class. Even if only a few samples are registered for each class, high-precision attribution can be achieved, which is very friendly to the maintenance cost and scalability of the database.

Samples	Real	T2VZ	MS	VC2	Pika	SVD	MuseV	Mora	CogV	Avg.
1	90.0	100.0	84.6	77.8	76.2	60.4	94.0	96.9	83.9	84.8
5	86.8	100.0	86.4	79.0	81.2	87.7	83.7	97.2	78.7	86.7
10	89.5	100.0	85.8	76.8	76.4	89.6	89.0	96.3	87.1	87.8
20	87.2	100.0	81.7	83.7	72.4	84.6	86.0	97.7	88.6	86.9
30	84.2	100.0	81.0	84.6	80.6	89.5	77.6	95.1	89.4	87.0
50	81.0	99.9	80.4	79.2	71.7	80.9	89.4	96.2	95.6	86.0

Table 7: **Analysis of registered samples per class.** We evaluate the sensitivity of our approach to the number of registered samples per class.