International Journal of Digital Crime and Forensics
Volume 18 • Issue 1 • January-December 2026

Vision Forgery Trace Enhanced VLMs for Generalized AIGC Video Detection

Lihua Wang
http://orcid.org/0009-0000-6151-2200
Nanjing University of Posts and Telecommunications, China & China Unicom Digital Technology Co., Ltd., China

Pengfei Pei
China Unicom Western Innovation Research Institute, China

Yiran He
Institute of Information Engineering, Chinese Academy of Sciences, China

Zihuan Huang
Xi'an Jiaotong University, China

Shuai Hu
Xi'an Jiaotong University, China & China Unicom Western Innovation Research Institute, China
Received: November 9th, 2025 | **Accepted:** February 23rd, 2026

ABSTRACT

Large vision language models (VLMs) show strong open-world generalization but degrade at domain-specific tasks, while traditional small forensic models perform well on in-distribution datasets yet lack cross-distribution generalization and language-based interpretability. To address this gap, the authors propose a vision forgery trace (VFT)-VLM framework, which incorporates forensic features into a VLM without sacrificing its general reasoning ability. Specifically, a lightweight VFT extraction module learns to encode texture anomalies, edge incoherence, pixel artifacts, and frequency-domain deviations. The traces are incorporated into the InternVL2-8B backbone via low rank adaptation fine-tuning, achieving alignment between visual evidence and textual explanations. Across 14 diverse artificial intelligence-generated content benchmark datasets, VFT-VLM outperforms VLM-based large-scale models and achieves comparable or superior performance to relevant traditional small-scale models. Ablation studies confirm both VFT extraction and low rank adaptation fine-tuning are critical to the performance gains.

KEYWORDS

Artificial Intelligence-Generated Content Video Detection, Low Rank Adaptation Tuning, Digital Forensics, VLM

INTRODUCTION

With the rapid progression of artificial intelligence-generated content (AIGC) technologies such as diffusion models and vision-language models (VLMs; Chen, Zhang, et al., 2024; Liu, Cun, et al., 2024; Wang, Chen, Ma, et al., 2024), generating high-fidelity images and videos has become increasingly feasible. On short-video sharing platforms like TikTok and YouTube, such content can be disseminated to a global audience within minutes. However, these technological advancements have also introduced notable security and societal risks. For instance, forged content targeting political figures or critical events may be exploited to manipulate public opinion, thereby posing severe threats

DOI: 10.4018/IJDCF.403419

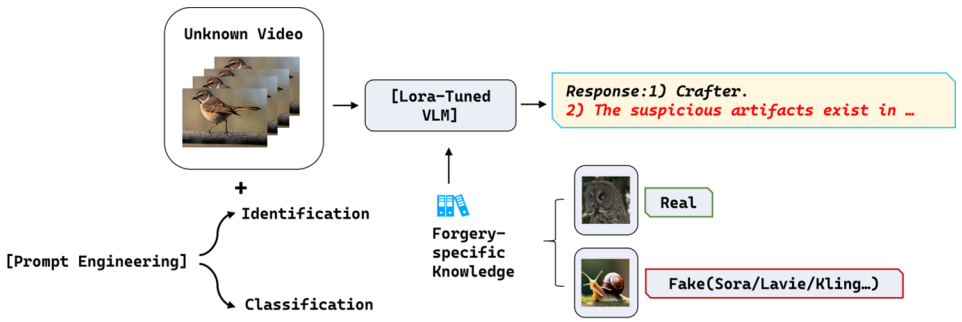
to individual privacy, social stability, and national security (Fu, 2024; S. Jia et al., 2024; Pei et al., 2023). Against this backdrop, there is an urgent demand for an efficient, robust, and generalized AIGC video detection system to safeguard the authenticity of digital content (Pei et al., 2025; K. Zhang, Peng, et al., 2024).

Existing detection methods predominantly consist of traditional small-scale models that rely on handcrafted features. Approaches including edge artifact trace analysis (Dong et al., 2023), Joint Photographic Experts Group (JPEG) compression estimation (Niu et al., 2021), noise distribution analysis (Pei et al., 2025), and self-attention for global feature capture (Yan et al., 2023) have been applied to image and video forgery detection. These methods attain an exceptionally high performance on specific datasets but suffer from limited interpretability and generalizability. This limitation becomes particularly prominent when confronted with broader and more diverse datasets, including real-world imbalanced data distributions (Dong et al., 2023; Liu, Li, et al., 2024; K. Zhang, Peng, et al., 2024).

Recent advancements in VLMs have provided inspiration for improving generalization and interpretability in AIGC detection. With their massive parameters, VLMs perform excellently in general tasks and exhibit robust text output capabilities, which facilitate adaptation to diverse downstream tasks and result interpretation. For instance, FKA-Owl (Liu, Li, et al., 2024), the first VLM-based image detection method, incorporates an extra image processing branch to enhance detection performance, whereas S. Jia et al. (2024) further investigated the application of ChatGPT in deepfake detection. Nevertheless, the untrained paradigm of such models leads to only moderate accuracy in AIGC forensics. Despite these endeavors, VLM-based methods for AIGC video detection remain largely underexplored.

Figure 1 shows the workflow of VFT-VLM. First, prompt engineering breaks down the task into identification and classification to prepare question-and-answer training data. During the training phase, the designed vision-forgery-specific knowledge is embedded into the VLM, and low rank adaptation (LoRA) is used to fine-tune the VLM. This enables the model to determine whether videos are real or fake and output textual explanations for the detected AIGC methods.

Figure 1. Workflow



To address these challenges, we propose incorporating vision forgery trace (VFT) into VLM, creating a novel framework tailored to enhance the forgery detection capabilities of VLMs. As illustrated in Figure 1, we adopt the pre-trained InternVL2-8B model as the backbone. Specifically, we incorporate a VFT extraction (VFTE) module to capture key forgery trace features. This integration of VLMs with vision-forgery-specific knowledge features derived from traditional methods is particularly crucial for advancing AIGC detection. In the model training phase, to align visual and textual representations and better model the semantic relationships between forged visual