

# VIDS-guard: A novel forensics-aware multi-stream transformer framework for robust deepfake video detection

Sami Alanazi<sup>\*</sup> , Seemal Asif<sup>†</sup> 

Faculty of Engineering and Applied Sciences, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom

## ARTICLE INFO

### Keywords:

Deepfake detection  
Video forensics  
Temporal modeling  
Frequency embedding  
Transformer architecture  
Multimedia security

## ABSTRACT

The proliferation of highly realistic deepfake videos poses a growing threat to digital trust, underscoring the need for detectors that remain reliable across diverse manipulation types and capture conditions. This paper introduces VIDS-Guard (Video Integrity Deepfake Shield), a novel forensics-aware multi-stream transformer framework that integrates spatial, frequency, and temporal cues within a unified architecture. Unlike conventional convolutional or transformer-based detectors that rely primarily on semantic consistency, VIDS-Guard embeds forensic inductive biases through Spatial Rich Model (SRM) residual filtering, YCbCr color-space decomposition, and Fast Fourier Transform (FFT) spectral embeddings to expose subtle manipulation artifacts. A temporal transformer encoder with attention pooling further models cross-frame inconsistencies, enabling robust video-level predictions.

Extensive experiments conducted with six benchmark models—Xception, ResNet50, MobileNetV3-Large, SlowFast, ViViT, and TimeSformer—demonstrate that VIDS-Guard achieves superior generalization and balanced detection performance across validation, test, and unseen datasets, attaining the highest accuracy and Macro-F1 scores under domain shift. These findings establish VIDS-Guard as a state-of-the-art forensic framework for trustworthy multimedia authentication and emphasize the importance of incorporating forensic priors to ensure sustainable robustness in deepfake video detection.

## 1. Introduction

The rapid progress of generative video synthesis has enabled the creation of highly realistic and virtually undetectable deepfakes, raising critical concerns for digital trust, social stability, and forensic integrity (Tolosana et al., 2020; Verdoliva, 2020). As generative models advance from GAN-based pipelines to diffusion-driven and hybrid frameworks, the synthetic traces they leave become increasingly subtle and less correlated with traditional visual artifacts. These developments challenge forensic practitioners, social platforms, and policymakers tasked with distinguishing synthetic from authentic media at scale. The consequences of misclassification are significant, ranging from political manipulation and reputational harm to the submission of falsified media as legal evidence.

### 1.1. State-of-the-art paradigms and the generalization challenge

Deepfake detection research has evolved along three dominant paradigms:

- **Frame-based convolutional detectors** leverage spatial semantics but treat deepfakes as standard image classification problems, underutilizing artifact-level cues and failing when manipulations span multiple frames (Qian et al., 2020; Rossler et al., 2019).
- **Temporal models** such as recurrent and transformer-based architectures exploit sequential dynamics but often rely only on RGB representations, ignoring spectral and chroma inconsistencies introduced during synthetic generation (Bertasius et al., 2021; Güera & Delp, 2018).
- **Physiological signal-based detectors** (e.g., rPPG-based systems) exploit biological regularities but remain vulnerable to illumination, pose variation, and compression distortions, limiting deployment reliability.

Across these paradigms, a recurring failure mode persists: detectors perform well on in-distribution benchmark datasets yet degrade sharply when exposed to unseen manipulations, compression levels, or acquisition pipelines. This limitation highlights a fundamental research gap in generalization beyond controlled training distributions.

<sup>\*</sup> Corresponding author.

E-mail addresses: [sami.alanazi@cranfield.ac.uk](mailto:sami.alanazi@cranfield.ac.uk) (S. Alanazi), [s.asif@cranfield.ac.uk](mailto:s.asif@cranfield.ac.uk) (S. Asif).

<https://doi.org/10.1016/j.iswa.2026.200664>

Received 29 January 2026; Received in revised form 6 March 2026; Accepted 8 April 2026

Available online 9 April 2026

2667-3053/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1.2. Research gap and motivation

Current deepfake detectors predominantly learn semantic or structural representations, but synthetic media generation alters the signal space in ways not captured by semantics alone. Frequency distortions, micro-boundary inconsistencies, chroma leakage, and motion–texture mismatch encode forensic evidence that persists even when visual realism is high. However, existing methods fail to fully integrate these cues:

- Existing spatial CNNs ignore many of these forensic artifacts, treating detection as conventional recognition rather than digital forensics.
- Temporal models capture motion patterns but lack mechanisms to attend to artifact-specific cues, especially under compression or re-encoding.
- Physiological-based approaches, while promising, are noise-sensitive and require integration into architectures robust enough for real-world forensic conditions.

Motivated by these limitations, we posit that **forensics-aware inductive biases** should be explicitly embedded into the architecture rather than expected to emerge implicitly from supervised learning. This paper introduces VIDS-Guard to bridge the generalization gap by systematically fusing multiple complementary forensic signals (spectral, residual, and temporal) within a robust transformer architecture.

## 1.3. Our approach: the VIDS-guard framework

To address these challenges, we introduce VIDS-Guard, a multi-stream transformer-based framework that fuses spatial, frequency, and temporal forensic cues within a unified architecture. The design integrates:

- **SRM-based residual filtering** to amplify high-frequency manipulation traces.
- **YCbCr chroma decomposition** to expose cross-channel inconsistencies.
- **FFT-based frequency embeddings** to reveal periodic and compression-related artifacts.
- **A temporal transformer encoder with attention pooling** to aggregate motion and cross-frame irregularities.

This configuration allows the network to jointly reason over semantic content and forensic distortions that are diagnostically relevant but visually subtle, resulting in improved generalization across unseen manipulations and acquisition pipelines.

## 1.4. Contributions

The main contributions of this work are:

1. A novel forensics-aware multi-stream architecture that embeds SRM, YCbCr, and FFT inductive biases into a temporal transformer for deepfake video detection.
2. A Unified Forensic Representation that exploits high-frequency residuals (SRM), chroma-luma inconsistencies (YCbCr), and periodic spectral artifacts (FFT) to enhance robustness under compression and manipulation diversity.
3. Extensive comparative evaluation against six benchmark architectures under matched conditions across validation, test, and unseen datasets.
4. Ablation analysis demonstrating the complementary impact of each forensic stream and the critical role of temporal modeling.
5. The public release of the full VIDS-Guard v1.0 dataset and a standardized evaluation pipeline to promote reproducible, artifact-centric research.

VIDS-Guard advances deepfake video detection by demonstrating that explicit forensic modeling—not semantic recognition alone—is essential for sustainable generalization in real-world multimedia security and forensic applications. The remainder of this paper is organized as follows: [Section 2](#) reviews related work. [Section 3](#) details the proposed VIDS-Guard architecture. [Section 4](#) presents the experimental setup and results, and [Section 5](#) concludes the paper.

## 2. Related works

### 2.1. Frame-based spatial detectors

Early deepfake detection efforts focused on single-frame analysis. [Rossler et al. \(2019\)](#) introduced FaceForensics++, establishing benchmark CNN baselines for manipulated images. Xception ([Chollet, 2017](#)) was later adopted as a strong spatial backbone due to its depthwise separable convolutions. Frequency-aware spatial approaches such as F3-Net ([Qian et al., 2020](#)) and SFIAD ([Kou et al., 2025](#)) enhanced robustness by leveraging localized spectral statistics. While these detectors demonstrate strong performance on curated benchmarks, treating detection purely as a spatial classification task limits their ability to capture cross-frame inconsistencies. This results in performance degradation under real-world compression and unseen manipulation pipelines.

### 2.2. Temporal modeling of video artifacts

Given that deepfakes are generated frame-by-frame, modeling temporal inconsistencies is critical. CNN-RNN architectures ([Güera & Delp, 2018](#)) and spatiotemporal CNNs ([De Lima et al., 2020](#)) exploit sequential information, while 3D CNNs ([Ganiyusufoglu et al., 2020](#)) enhance motion reasoning and robustness. Transformer-based architectures, including TimeSformer ([Bertasius et al., 2021](#)) and ViViT ([Arnab et al., 2021](#)), further improve long-range modeling capabilities. However, temporal models relying solely on RGB representations may overlook subtle forensic cues—compression artifacts, spectrum perturbations, and chromatic leakage—that remain critical under domain shift.

### 2.3. Frequency and forensic priors

Forensic artifacts introduced during generative synthesis frequently manifest in the frequency domain. Early studies explored JPEG double-compression patterns for forgery detection ([Barni et al., 2017](#)). Recent works, including F3-Net ([Qian et al., 2020](#)), wavelet-based approaches ([Wang et al., 2023](#)), and FreqNet ([Tan et al., 2024](#)), demonstrate that spectral signals can survive post-processing and offer stronger cross-GAN generalization. However, these techniques primarily focus on single images and often lack temporal awareness, making them less effective for video scenarios where manipulation quality varies across frames.

### 2.4. Physiological and rPPG-based methods

Physiological cues represent an orthogonal direction in deepfake forensics. FakeCatcher ([Ciftci & Demir, 2020](#)) exploits subtle changes in blood volume using remote photoplethysmography (rPPG), followed by pulse-based methods such as DeepFakesON-Phys ([Hernandez-Ortega et al., 2020](#)) and subsequent refinements in temporal modeling ([Xu et al., 2023](#)). AVENUE ([Birla et al., 2025](#)) further improves robustness by integrating rPPG with a Temporal Convolutional Network (TCN).

Although these methods demonstrate promise, their accuracy can be disrupted by illumination fluctuations, pose variation, low-resolution sensors, and facial occlusions. Thus, while physiological cues are informative, they require integration within architectures that are specifically designed to mitigate noise sensitivity and multi-factor distortion.

## 2.5. Generalization and robustness across domains

Generalization to unseen datasets remains one of the most persistent challenges. Capsule-based learning (Nguyen et al., 2019), transfer learning (Lee et al., 2021), and interpretable transformer architectures (Zhao et al., 2023) have attempted to address domain shift. The DFDC benchmark (Dolhansky et al., 2020) highlighted that even competitive detectors experience sharp degradation when exposed to new manipulation families, codec variations, or acquisition pipelines. The literature converges on a key observation: robustness requires models to move beyond semantic content and explicitly capture forensic trace distributions.

## 2.6. Positioning of VIDS-guard

Building on the strengths and recognized limitations of prior research, **VIDS-Guard proposes a novel unified forensic framework that explicitly targets the gaps in cross-domain robustness and multi-artifact detection.**

VIDS-Guard systematically addresses the convergence of challenges by integrating four essential components:

- **Addressing Spatial Limitations:** Unlike spatial-only methods (Qian et al., 2020; Rossler et al., 2019), which fail under cross-frame inconsistencies, VIDS-Guard explicitly embeds forensic priors through the use of Steganalysis Rich Model (SRM), YCbCr streams, and Fast Fourier Transform (FFT) channels. This multi-stream input ensures artifact-specific features are prioritized over semantic content.
- **Enhancing Temporal Modeling:** Where purely temporal architectures (Arnab et al., 2021; Bertasius et al., 2021) rely predominantly on subtle RGB differences, VIDS-Guard fuses the sequential attention of a transformer architecture with the artifact-specific features derived from the frequency and color streams. This allows it to model sequential changes in forensic traces, rather than just general motion.
- **Extending Frequency Analysis:** While previous frequency-based techniques (Barni et al., 2017; Tan et al., 2024) typically focus on single images, our integration of frequency streams within the spatiotemporal transformer enables VIDS-Guard to capture how spectral signals and compression artifacts vary across the video timeline.
- **Improving Robustness over Physiological Cues:** Compared to rPPG-based methods (Birla et al., 2025; Ciftci & Demir, 2020), VIDS-Guard avoids exclusive dependence on noise-sensitive biological signals, which are easily disrupted by lighting and pose variation. Our approach maintains accuracy across diverse acquisition conditions while remaining flexible for future integration of refined physiological cues.

This deliberate balance—fusing sequential temporal modeling with multiple, artifact-aware forensic feature streams—enables VIDS-Guard to sustain high accuracy and superior robustness under real-world challenges, including high compression, codec variation, and unseen manipulation pipelines.

## 3. Methodology

### 3.1. Dataset

#### 3.1.1. Main deepfake video dataset

The primary dataset used in this study, referred to as the VIDS-Guard Dataset (v1.0) (Alanazi & Asif, 2025a, 2025b), comprises 26,975 videos, evenly balanced between real and fake content (13,487 real and 13,488 fake). The collection integrates material from eleven publicly available benchmarks, ensuring broad coverage of manipulation techniques including GAN-based full synthesis, identity swap, expression

reenactment, and codec-induced distortions.

- CelebV-HQ (Zhu et al., 2022) provides high-quality real videos with wide variation in pose, lighting, and facial expression.
- Identity swapping. Datasets like Celeb-DF (Li et al., 2020a, 2020b) and UADFV (Keshri, 2024) contain manipulated videos where the facial identity of the source is replaced with that of a target.
- Expression reenactment. FaceForensics++ (Rossler et al., 2019) and DeepfakeTIMIT (Korshunov & Marcel, 2018) focus on altering expressions or speech synchrony while retaining subject identity.
- Hybrid and compression artifacts. Large-scale benchmarks such as DFDC (Dolhansky et al., 2020) and DFD (Google & Jigsaw, 2019) introduce additional challenges with post-processing, recompression, and codec variations.

Real-world variability is preserved through diversity in lighting, pose, ethnicity, camera quality, and background conditions. Table 1 presents the distribution of real and fake videos across contributing datasets.

### Subsampling and Dataset Harmonization

To achieve the final balanced totals (13,487 real / 13,488 fake), we uniformly subsampled real videos from large real-only sources (e.g., CelebV-HQ, FakeAVCele) after deduplication and metadata harmonization.

Subsampling was performed at the video level using stratified random selection constrained by identity metadata when available. Specifically, for datasets providing subject labels (e.g., CelebV-HQ, Celeb-DF), identity-level grouping was first established to avoid overrepresentation of dominant individuals. Real-only datasets were then proportionally reduced to maintain class balance without altering the distribution of manipulation types in the fake class. No synthetic oversampling, duplication, or artificial augmentation was used to equalize class counts.

All preprocessing steps—including deduplication, renaming, and final per-dataset counts—are documented in the publicly released Zenodo README files accompanying VIDS-Guard Dataset (v1.0) (Parts 1 and 2).

### Split Policy and Identity-Based Partitioning

For datasets containing multiple videos per identity, subject-disjoint splitting was enforced. All videos associated with a given identity—real and manipulated—were assigned exclusively to a single split to prevent identity leakage across training and evaluation stages. This ensures that the model cannot implicitly memorize subject-specific facial characteristics.

For identity-swapping datasets (e.g., Celeb-DF, FaceForensics++), both the authentic and manipulated versions of a given subject were placed within the same partition. For datasets lacking explicit identity

**Table 1**  
Composition of the VIDS-guard dataset (v1.0).

| Dataset Name            | Real Videos | Fake Videos | Reference                 |
|-------------------------|-------------|-------------|---------------------------|
| CelebV-HQ               | 9,442       | 0           | Zhu et al. (2022)         |
| DeeperForensics-1.0     | 800         | 0           | Jiang et al. (2020)       |
| Celeb-DF v1             | 300         | 795         | Li et al. (2020a)         |
| Celeb-DF v2             | 392         | 3,669       | Li et al. (2020b)         |
| DeepFakeDetection (DFD) | 364         | 3,067       | Google & Jigsaw (2019)    |
| DFDC (Kaggle)           | 300         | 3,664       | Dolhansky et al. (2020)   |
| UADFV                   | 49          | 49          | Keshri (2024)             |
| DeepfakeTIMIT           | 0           | 318         | Korshunov & Marcel (2018) |
| Kaggle Selfie Dataset   | 40          | 40          | KUCEV (2023)              |
| FaceForensics++ (FF++)  | 1,000       | 1,886       | Rossler et al. (2019)     |
| FakeAVCele              | 800         | 0           | Khalid et al. (2021)      |
| Total                   | 13,487      | 13,488      | —                         |

metadata, filename clustering and metadata inspection were used to minimize cross-split overlap.

Validation data were used exclusively for model selection and early stopping. The internal test set remained fully held out for in-distribution evaluation.

### 3.1.2. External test set (unseen data)

To rigorously evaluate cross-dataset generalization, VIDS-Guard was assessed on an external unseen dataset (Version 1.0) containing 600 videos (300 real, 300 fake). This subset, distributed separately as Part 2 of the Zenodo release (Alanazi & Asif, 2025b), integrates material from five independent sources alongside an unseen subset of CelebV-HQ. None of these samples were exposed during training or validation, providing a stringent benchmark for robustness.

A key component of this external benchmark is the HiDF dataset (Chaewon et al., 2025), a recently introduced collection comprising over 62,000 images and 8,000 videos generated with commercial deepfake tools. HiDF content has been manually curated to be indistinguishable from authentic media by human evaluators, making it a challenging frontier for forensic detection systems. By incorporating HiDF samples, our evaluation captures failure modes that may arise in next-generation synthetic content.

Other sources include smaller Kaggle datasets (Steven, 2025; Unidata, 2025), the Anti-Spoofing Real Dataset (2023), and Real vs Fake Liveness Detection Dataset (2023) (Unique Data, 2023a, 2023b), each contributing distinct acquisition conditions. An unseen subset of CelebV-HQ (Zhu et al., 2022) further provides real high-quality control data. The diversity of these sources ensures that the external test set captures both low-resource manipulations and high-quality, human-indistinguishable forgeries. Table 2 summarizes the composition of the external test set.

The external dataset was kept strictly separate from the training pipeline and employed solely for final evaluation. By incorporating manipulations and acquisition conditions absent during model development, including human-indistinguishable forgeries from HiDF, the test set provides a rigorous measure of cross-dataset generalization and serves as a realistic proxy for deployment scenarios.

### 3.1.3. Ethics and compliance

All datasets incorporated into the VIDS-Guard Dataset (v1.0) and the external unseen evaluation set are publicly released research resources. Their use in this study strictly adheres to the respective licenses and terms of access. Only face data already available within these datasets were processed, and no new personal data were collected or generated. All experiments were conducted in compliance with institutional and data-sharing guidelines for research reproducibility and ethical AI development.

## 3.2. Pre-processing and clip sampling

Prior to model training, video frames were pre-processed to ensure

**Table 2**  
Composition of the External Test Set (Unseen Data, Version 1.0).

| Dataset Name                      | Real Videos | Fake Videos | Reference             |
|-----------------------------------|-------------|-------------|-----------------------|
| Deepfake Dataset                  | 6           | 32          | Steven (2025)         |
| Anti-Spoofing Real Dataset        | 6           | 0           | Unique Data (2023a)   |
| Real VS Fake – Liveness Detection | 11          | 0           | Unique Data (2023b)   |
| DeepFake Videos Dataset           | 0           | 1           | Unidata (2025)        |
| HiDF                              | 0           | 267         | Chaewon et al. (2025) |
| CelebV-HQ (Unseen Subset)         | 277         | 0           | Zhu et al. (2022)     |
| Total                             | 300         | 300         | —                     |

consistency and robustness against real-world degradations. Face regions were first extracted using the MTCNN detector (Zhang et al., 2016), followed by temporal tracking to maintain alignment across consecutive frames. From each video, fixed-length clips were constructed by uniformly sampling  $T$  frames with temporal jitter to capture natural variations; in this study, we set  $T = 8$ .

Each sampled frame was resized to  $224 \times 224$  pixels and normalized using the ImageNet mean and standard deviation. During training, a series of data augmentations was applied to simulate the degradations commonly introduced by online video distribution platforms. These included random resized cropping, horizontal flipping, JPEG re-compression with quality  $q \in [30, 90]$ , Gaussian blur, brightness/contrast jitter, and temporal perturbations through random frame duplication or omission.

The final input tensor shape is defined as:

$$X \in \mathbb{R}^{B \times T \times 3 \times 224 \times 224}$$

where  $B$  is the batch size,  $T$  is the number of frames per clip, and 3 is the RGB channels.

### 3.3. VIDS-guard: forensics-aware architecture and workflow

The proposed VIDS-Guard (Fig. 1) is an end-to-end deepfake video detector that explicitly integrates forensic cues into the backbone. Unlike conventional CNN/Transformer models that primarily capture semantic features, VIDS-Guard emphasizes manipulation-specific artifacts across spatial, frequency, and temporal dimensions.

Workflow. Input face clips are extracted, resized, normalized, and processed by four coordinated modules:

1. **Multi-branch forensic encoder (VGBackbone)** that produces per-frame forensic embeddings.
2. **Frequency embedding** via FFT-based spectral decomposition to preserve explicit spectral cues.
3. **Temporal modeling** with Transformer encoder layers to aggregate cross-frame inconsistencies (e.g., expression misalignment, blink irregularities, photometric flicker).
4. **Temporal attention pooling + binary classification**, which adaptively emphasizes frames carrying the strongest manipulation evidence to yield a robust video-level prediction.

This forensics-aware workflow prevents over-reliance on semantics and instead exploits manipulation-specific traces across multiple domains, enabling strong generalization to unseen datasets. Table 3 outlines the main components, output dimensions, parameter counts, and roles within the pipeline. The VGBackbone (three-stream spatial encoder) contains approximately 15M parameters, distributed across the SRM, YCbCr, and frequency-proxy branches using depthwise-separable convolutional blocks to maintain computational efficiency. The FFT embedding module contributes  $\sim 0.1$ M parameters. The 2-layer Transformer encoder (8 heads, hidden dimension 640) contains  $\sim 9.8$ M parameters, while attention pooling and the classification head contribute  $\sim 0.4$ M and  $\sim 0.6$ M parameters, respectively. The total number of trainable parameters in VIDS-Guard is approximately 25.9M.

#### 3.3.1. Multi-branch spatial encoder (VGBackbone)

The VGBackbone serves as the spatial encoder and is specifically engineered to capture forensic cues overlooked by conventional CNNs. It consists of three parallel feature streams:

- **Residual stream (SRM):** Applies 30 fixed steganalysis-inspired high-pass SRM filters (as commonly adopted in digital forensics literature) followed by depthwise-separable convolutions to emphasize blending seams and high-frequency artifacts. These filters

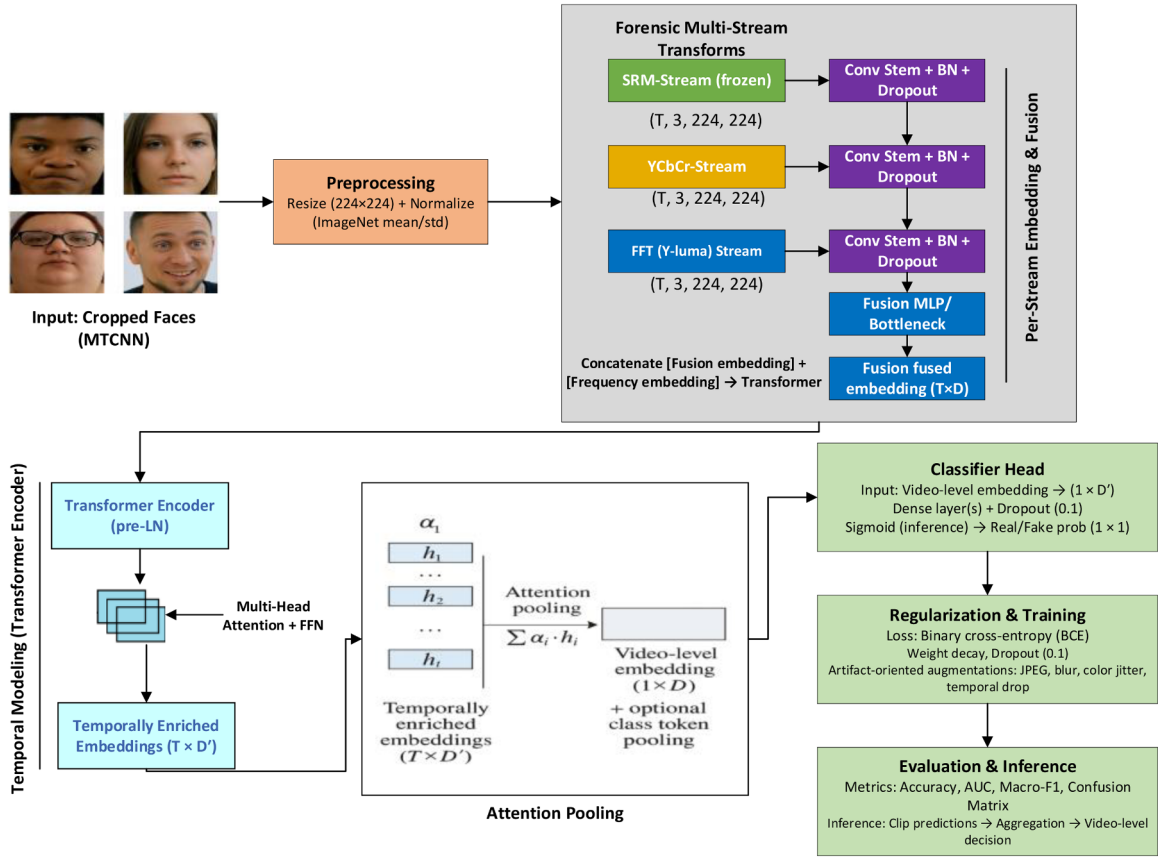


Fig. 1. VIDS-Guard architecture integrating multi-stream forensic cues (SRM, YCbCr, FFT) with temporal transformer and attention pooling for video-level deepfake classification.

**Table 3**  
Summary of the VIDS-Guard pipeline.

| Stage               | Component                       | Output Shape | Parameters | Role                                          |
|---------------------|---------------------------------|--------------|------------|-----------------------------------------------|
| Input               | Face clip (B, T, 3, 224, 224)   | –            | –          | Preprocessed face crops                       |
| Backbone            | VGBBackbone (3-stream CNN)      | (B × T, 512) | ~15M       | SRM/YCbCr/frequency-proxy fusion              |
| Frequency           | FFT band decomposition          | (B, T, 128)  | ~0.1M      | Preserves spectral cues                       |
| Fusion              | Concatenation                   | (B, T, 640)  | –          | Concatenates spatial and spectral embeddings. |
| Temporal            | Transformer (2 layers, 8 heads) | (B, T, 640)  | ~9.8M      | Models cross-frame dependencies.              |
| Pooling             | Attention pooling               | (B, 640)     | ~0.4M      | Emphasize informative frames                  |
| Classification Head | Dense + Sigmoid                 | (B, 1)       | ~0.6M      | Outputs real/fake probability.                |

are non-trainable and serve as forensic priors to enhance residual inconsistencies before convolutional refinement.

- **Chroma-luminance stream (YCbCr):** Converts frames into Y, Cb, and Cr channels to capture chroma inconsistencies and cross-channel leakage unique to GAN-generated frames.
- **Frequency-proxy convolution stream:** Incorporates BlurPool anti-aliasing and depthwise convolutions to approximate DCT-like local spectral patterns associated with compression.

Outputs from these three streams are fused via a bottleneck MLP and projected into a 512-dimensional embedding per frame, balancing semantic and forensic features.

Furthermore, the three streams are fused into a unified embedding per frame. The fusion is defined as:

$$z_{frame} = \text{GAP}([f_{SRM} || f_{YCbCr} || f_{DCT}]) \in \mathbb{R}^{512}$$

Where  $f_{SRM}$ ,  $f_{YCbCr}$ ,  $f_{DCT}$  denote branch-specific features,  $[ \cdot || \cdot ]$  indicates concatenation, and GAP is global average pooling.

### 3.3.2. Frequency embedding module

Deepfake manipulations often leave detectable spectral signatures. To leverage this, a 2D FFT is computed over each frame, and the magnitude spectrum is partitioned into  $B = 8$  radial frequency bands. For each band, normalized energy statistics are computed and projected into a 128-dimensional learnable embedding. The FFT embedding module contains approximately 0.1M trainable parameters. The resulting 128-D spectral embedding is concatenated with the 512-D spatial features prior to temporal modeling.

This pathway preserves spectral information that would otherwise be suppressed in conventional convolutional pipelines, ensuring that frequency cues remain accessible to the temporal encoder.

Formally, the FFT magnitude spectrum  $F(u, v)$  is partitioned into  $B$  radial bands. The band energy is computed as:

$$e_b = \frac{1}{|R_b|} \sum_{(u,v) \in R_b} |F(u, v)|, b = 1, \dots, B$$

A learnable embedding is then obtained via:

$$z_{freq} = w^T E \in \mathbb{R}^{128}$$

where  $E$  is the band-embedding matrix and  $W$  are learnable weights.

### 3.3.3. Temporal transformer encoder

The concatenated spatio-frequency features are processed by a 2-layer Transformer encoder with 8 heads. This module captures both short-term inconsistencies (frame-to-frame flicker) and long-term anomalies (illumination or expression drift). Stability is ensured through GELU activations, residual connections, and layer normalization.

The encoder updates hidden states as:

$$H^{(l)} = \text{MHSA}(\text{LN}(H^{(l-1)})) + H^{(l-1)}, H^{(l)} = \text{FFN}(\text{LN}(H^{(l)})) + H^{(l)}$$

where MHSA is multi-head self-attention, LN is layer normalization, and FFN is the feed-forward block.

### 3.3.4. Attention pooling and classification

The aggregated video-level representation is obtained through a learnable attention mechanism that adaptively weights frame embeddings:

$$Z = \sum_{t=1}^T a_t h_t, \quad a_t = \frac{\exp(w^T h_t)}{\sum_{j=1}^T \exp(w^T h_j)}$$

Where  $h_t$  is the hidden state for frame  $t$ , and  $a_t$  is its attention weight normalized across all frames. This formulation enables the network to prioritize frames carrying stronger forensic cues while down-weighting less informative ones. The pooled embedding  $Z$  is passed through a fully connected layer with sigmoid activation to yield the final probability of a video being real or fake.

### 3.4. Inconsistency detection mechanism

As illustrated in Fig. 2, VIDS-Guard explicitly targets temporal and photometric inconsistencies that are characteristic of deepfake videos. The multi-stream extractor captures spatial forensic cues, which are then aggregated temporally to reveal:

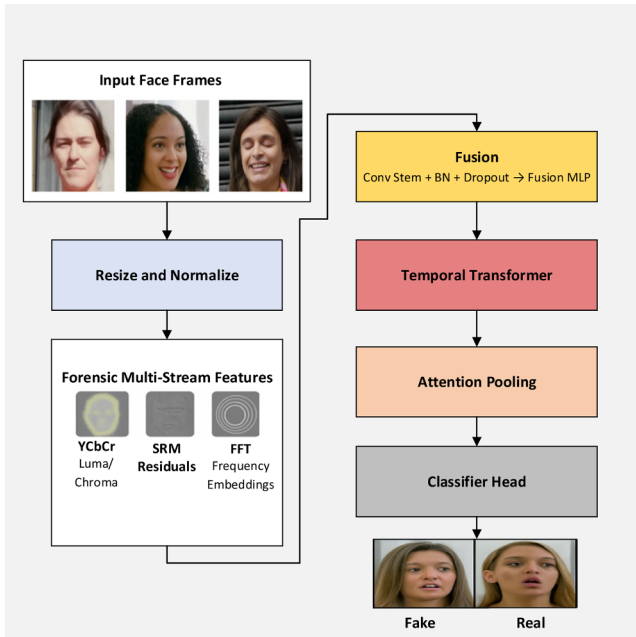


Fig. 2. VIDS-Guard detecting temporal and photometric inconsistencies in deepfake videos.

- **Facial dynamics anomalies:** expression mismatches, abnormal blinking, lip-sync drift, boundary flicker.
- **Lighting and photometric irregularities:** inconsistent specular highlights, relighting flicker, geometry distortions from compression.

The temporal transformer identifies these cross-frame inconsistencies, while attention pooling prioritizes frames that contain the strongest manipulation evidence. This mechanism ensures the detector focuses on artifact-driven irregularities rather than semantic content, enhancing robustness to unseen manipulations.

### 3.5. Training procedure

VIDS-Guard is trained with AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of  $2 \times 10^{-4}$  and weight decay of  $5 \times 10^{-2}$ . A cosine learning rate schedule with warm-up was adopted to stabilize early optimization and promote smooth convergence. Training was performed using a batch size of 8 on  $224 \times 224$  face crops extracted from video frames.

Model convergence was monitored using validation AUC and Macro-F1 as performance indicators. As illustrated in Fig. 3 (VIDS-Guard Training Loss and Validation AUC), training loss steadily decreased from approximately 0.356 at epoch 0 to 0.154 at epoch 11, while validation AUC increased rapidly after the initial epochs and stabilized in the 0.95–0.97 range from epoch 5 onward, peaking at approximately 0.97 around epoch 9. This behavior indicates stable optimization without evidence of overfitting.

The checkpoint at epoch 11 provided the most favorable trade-off between validation AUC (0.9697) and validation accuracy (0.9006) and was therefore selected for all subsequent evaluations.

To enhance robustness to real-world video degradation, we employed data augmentations including random frame dropping, temporal jitter, JPEG compression with quality factors in the range 30–90, Gaussian blur, and brightness–contrast perturbations. Automatic mixed precision (AMP) was enabled to improve computational efficiency and memory utilization.

### 3.6. Evaluation protocol

The trained models were evaluated using standard metrics, including Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), and Macro-F1, to capture both overall and class-balanced performance. Confusion matrices were additionally reported to visualize prediction distributions across real and fake categories.

Unless otherwise stated, classification thresholds were fixed at 0.5. Evaluation was conducted in two stages: (i) on the internal test set to measure in-distribution performance, and (ii) on an external unseen dataset to benchmark cross-dataset generalization.

### 3.7. Novelty and architectural relation to existing video deepfake detectors

Unlike conventional baselines such as Xception, EfficientNet, or Swin Transformer—which rely primarily on spatial features or generic temporal modules—VIDS-Guard integrates three distinctive technical contributions: (i) a custom forensic backbone (VGBBackbone) incorporating residual, chroma, and frequency streams; (ii) an explicit FFT-based frequency embedding preserved throughout the entire pipeline; and (iii) a temporal transformer with attention pooling for the selective aggregation of discriminative frames.

To our knowledge, VIDS-Guard is among the first end-to-end video-level frameworks to explicitly integrate residual (SRM), chroma (YCbCr), and FFT-based spectral priors within a unified temporal transformer architecture specifically optimized for forensic artifact extraction rather than generic action recognition. Table 4 provides a

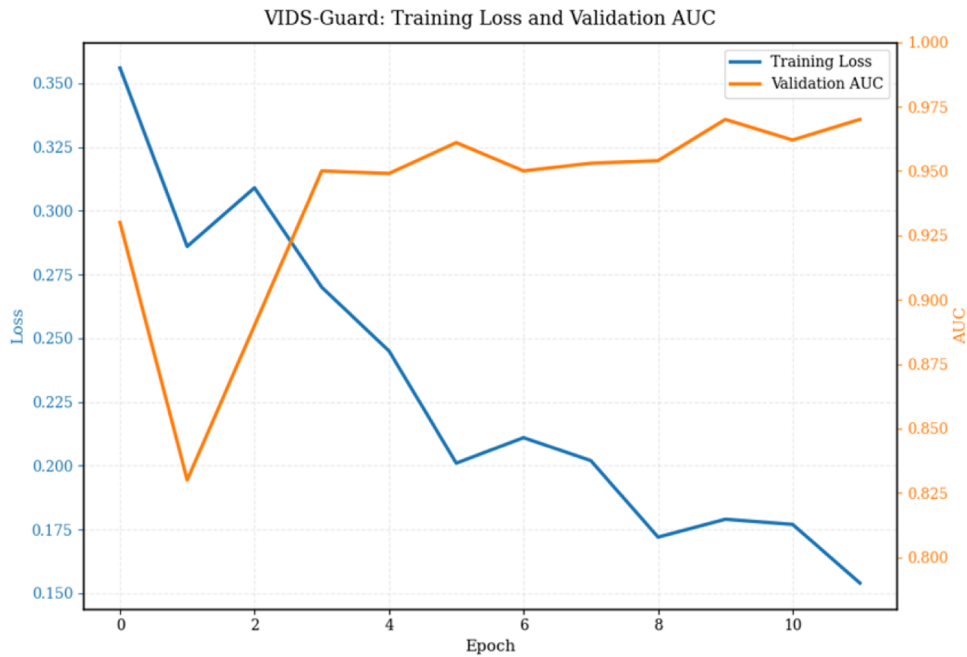


Fig. 3. VIDS-guard training loss and validation AUC.

Table 4

Technical properties and forensic capabilities of representative spatio-temporal architectures.

| Category           | Model             | Input Modality        | Temporal Strategy        | Key Characteristics                                          |
|--------------------|-------------------|-----------------------|--------------------------|--------------------------------------------------------------|
| <b>Image-based</b> | Xception / ResNet | RGB (Spatial)         | N/A                      | High efficiency; lacks video-level context                   |
| <b>Efficient</b>   | MobileNetV3       | RGB (Spatial)         | N/A                      | Edge-optimized; low capacity for complex forgeries           |
| <b>Video-SOTA</b>  | SlowFast / ViViT  | RGB (Motion)          | 3D Conv / Attention      | Captures long-range motion; ignores spectral traces          |
| <b>Proposed</b>    | <b>VIDS-Guard</b> | <b>RGB+FFT+Chroma</b> | <b>Attention Pooling</b> | <b>Multi-stream forensic awareness; adaptive aggregation</b> |

comparative qualitative analysis of VIDS-Guard against representative state-of-the-art (SOTA) architectures, including TimeSformer, ViViT, and SlowFast. Detailed quantitative performance metrics across validation, test, and cross-dataset benchmarks are further analyzed in Section 4.

## 4. Experimental results

### 4.1. Implementation setup

All experiments were conducted on the Delta 2 HPC cluster equipped with NVIDIA A100 GPUs (40 GB memory). Models were implemented in PyTorch (Paszke et al., 2019) with automatic mixed-precision (AMP) to optimize computational efficiency. Training used the AdamW optimizer (learning rate  $2 \times 10^{-4}$ , weight decay  $5 \times 10^{-2}$ ) with cosine scheduling and warmup. To improve robustness, augmentations included temporal jitter, frame drops, JPEG compression, Gaussian blur, and brightness/contrast variations.

Under this configuration, training required approximately 40 minutes per epoch with batch size 8 and temporal window  $T=8$  on a single NVIDIA A100 GPU.

Performance was evaluated on three splits: (i) validation set (10%), (ii) internal test set (10%), and (iii) an external unseen dataset of 600 videos from independent sources. Metrics included Accuracy, Precision, Recall, per-class F1-score, Macro-F1, and AUC. Confusion matrices were used to analyze error patterns and identify class bias.

The validation and test splits for VIDS-Guard contained 2,697 and 2,698 samples, respectively, while the baseline models were evaluated on 2,690 samples per split. The 8-sample discrepancy originates from face extraction failures during preprocessing, where a small number of

videos did not yield stable face crops required for consistent temporal clip construction. These samples were excluded prior to training and evaluation to maintain uniform input dimensionality across models. Importantly, the removed videos were evenly distributed across real and fake classes, preserving class balance and preventing bias in comparative evaluation.

### 4.2. VIDS-guard performance

The proposed VIDS-Guard model exhibited stable convergence and strong generalization across validation and test sets. As shown in Fig. 3 (Training Loss and Validation AUC), the training loss steadily decreased from 0.356 at epoch 0 to 0.154 at epoch 11. Validation AUC rose rapidly during early epochs and stabilized above 0.96, peaking at 0.97.

This stable convergence confirms that VIDS-Guard effectively captured discriminative forensic cues without overfitting.

On the validation split, the confusion matrix (Fig. 4) shows balanced detection of real and fake videos, with a Macro-F1 of 0.904, Precision range 0.892–0.916, and Recall range 0.881–0.925. No class imbalance was observed, contrasting with transformer-only baselines such as ViViT and TimeSformer.

On the test split, the confusion matrix (Fig. 5) reports 1,178 true negatives, 1,281 true positives, 141 false positives, and 98 false negatives, yielding Accuracy = 0.911 and Macro-F1 = 0.911. The ROC curve (Fig. 6) highlights an AUC of 0.976, demonstrating strong separability between real and fake classes.

On the unseen dataset (600 videos), VIDS-Guard maintained robust generalization under domain shift, achieving 274 true negatives, 266 true positives, 26 false positives, and 34 false negatives (Fig. 7). This corresponds to Accuracy = 0.900 and Macro-F1 = 0.900, confirming

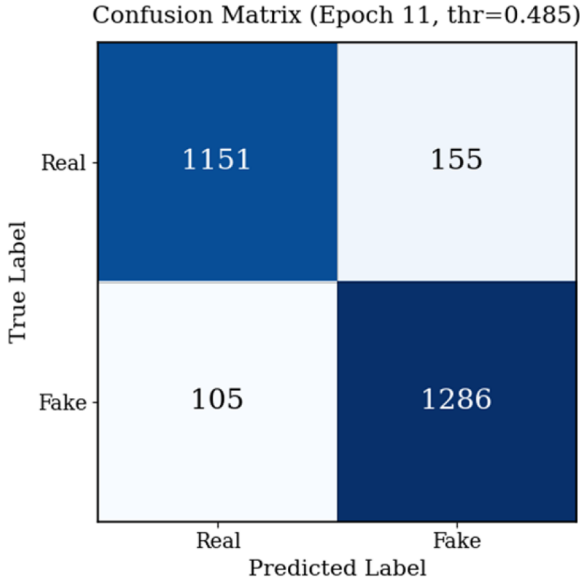


Fig. 4. VIDS-guard validation confusion matrix.

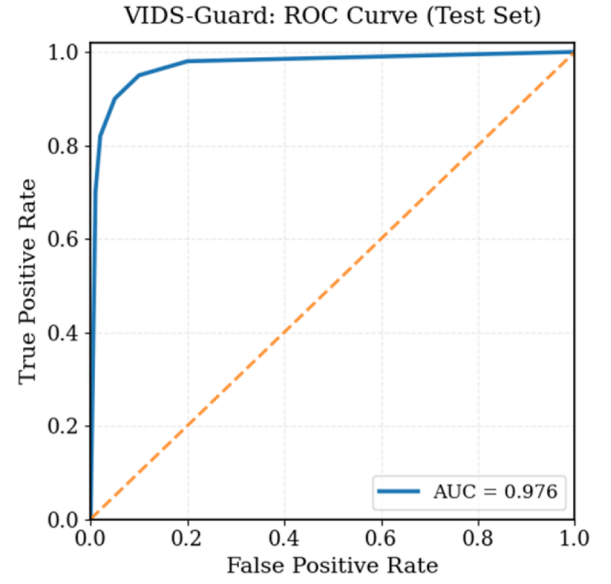


Fig. 6. VIDS-Guard ROC Curve on the Test Set.

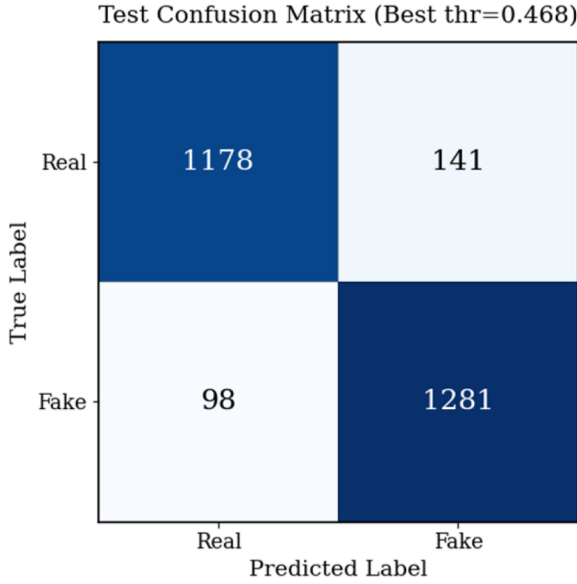


Fig. 5. VIDS-guard confusion matrix on the test set.

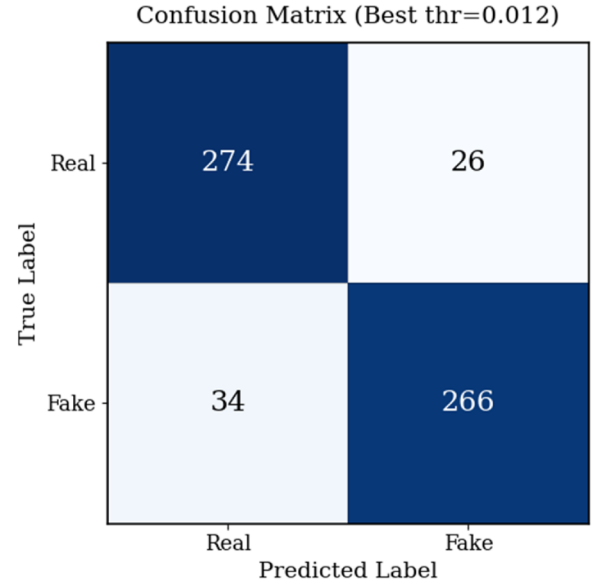


Fig. 7. VIDS-guard confusion matrix on the unseen dataset.

that VIDS-Guard generalizes effectively beyond the training distribution, outperforming competing architectures that degrade on unseen domains.

#### 4.3. Error case analysis

To further delineate the operational boundaries of VIDS-Guard, we conducted a qualitative analysis of misclassified samples from the unseen dataset. Figs. 8 and 9 illustrate two representative failure modes occurring under challenging forensic conditions.

- **Case 1 (False Negative: Subtle Manipulation)** Fig. 8 depicts a fake video (f445\_f01377) misclassified as real with a confidence score of 0.010. In this instance, the manipulation is visually seamless, characterized by high temporal stability in lighting and sharp spatial boundaries. These characteristics effectively suppress the typical GAN-generated artifacts that the SRM and FFT streams are designed to extract. Consequently, the model is forced to rely on semantic

consistency, which remains intact in high-fidelity deepfakes. This case highlights a persistent challenge in forensic detection: high-quality generative models can narrow the "forensic signal gap" by minimizing frequency-domain anomalies.

- **Case 2 (False Positive: Degraded Authentic Content)** Conversely, Fig. 9 shows a real video (qLq6GizMPu4\_1001\_1381) erroneously flagged as a forgery (confidence 0.029). This error stems from aggressive video compression and motion-induced blur. The resulting blocky textures and chroma subsampling artifacts in the YCbCr and SRM streams closely mimic the noise patterns typical of low-quality deepfakes. This suggests that the model occasionally struggles to differentiate between "intentional" manipulation noise and "accidental" transmission noise.

#### Insights and Future Directions

These failure cases reveal that VIDS-Guard remains vulnerable at two extremes: near-perfect fakes with negligible forensic noise, and heavily degraded authentic content. To address these limitations, future

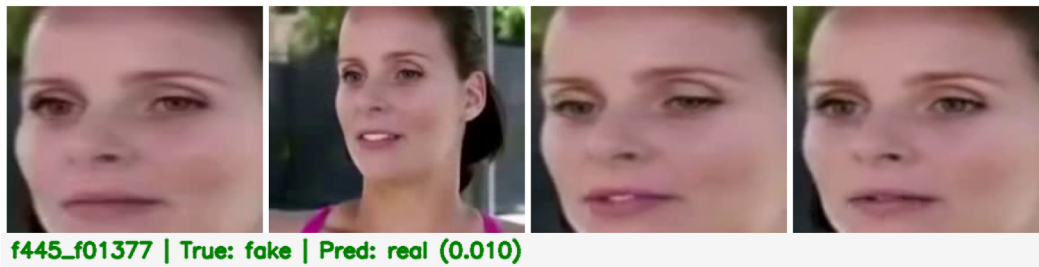


Fig. 8. False Negative – Fake video (f445\_f01377) misclassified as real (confidence 0.010). Subtle manipulations suppressed typical forensic cues.



Fig. 9. False Positive – Real video (qLq6GizMPu4\_1001\_1381) misclassified as fake (confidence 0.029). Compression and motion artifacts mimicked deepfake traces.

iterations of this work will explore:

1. Facial Dynamics Modeling: Integrating physiological signals (e.g., blinking patterns or pulse-based rPPG) to detect fakes that are spatially perfect but biologically inconsistent.
2. Compression-Aware Training: Implementing adversarial training on degraded samples to enhance the model's robustness against compression artifacts, thereby reducing the false positive rate on low-resolution content.

#### 4.4. Comparative evaluation with baseline models

To rigorously assess robustness and generalization capability, VIDS-Guard was systematically compared against six representative state-of-the-art deepfake detection architectures: Xception (Chollet, 2017), ResNet50 (He et al., 2015), MobileNetV3-Large (Howard et al., 2019), SlowFast (Feichtenhofer et al., 2019), ViViT (Arnab et al., 2021), and TimeSformer (Bertasius et al., 2021).

All models were trained and evaluated under strictly matched experimental conditions, employing identical preprocessing, augmentation strategies, and data splits (10% validation, 10% test), in addition to a 600-video external unseen dataset, ensuring a fair and controlled comparison.

##### Structured evaluation design

Rather than organizing results sequentially by model, the evaluation is structured into five analytical clusters, each aligned with a specific evaluation objective. This design choice dramatically improves interpretability by enabling direct comparison across critical performance dimensions. These dimensions include optimization stability, generalization, discrimination, and robustness, thereby avoiding repetitive per-model narratives.

Table 5 summarizes the five analytical clusters, outlining the specific objective, the key metrics used, and the corresponding performance dimension that is assessed in the subsequent Fig.s. This structured approach ensures every critical aspect of VIDS-Guard's performance is systematically compared against the baseline models.

##### Cluster 1: Training Dynamics and Convergence Stability (Training Loss and Validation AUC)

This cluster examines optimization behavior, convergence stability, and early overfitting tendencies across all evaluated models. Analyzing

Table 5

Summary of the Five Analytical Clusters for Comparative Model Evaluation.

| Cluster | Evaluation Objective          | Key Metrics / Visualizations      | Performance Dimension              |
|---------|-------------------------------|-----------------------------------|------------------------------------|
| 1       | Training Dynamics             | Training Loss, Validation AUC     | Optimization Stability             |
| 2       | In-Distribution Validation    | Validation Confusion Matrices     | Model Selection / Class Balance    |
| 3       | Test Set Generalization       | Test Confusion Matrices           | Performance Consistency            |
| 4       | Discrimination Ability        | Test ROC Curves                   | Threshold-Independent Separability |
| 5       | Robustness Under Domain Shift | Unseen Dataset Confusion Matrices | Real-World Applicability           |

these dynamics provides insight into how robustly each architecture learns the necessary feature representations.

Fig. 10 presents the training loss and validation AUC trajectories under identical optimization settings. The curves facilitate a direct comparison of learning dynamics, highlighting differences in convergence speed, stability, and susceptibility to performance degradation during training.

##### Cluster 2: In-Distribution Validation Performance (Validation Confusion Matrices)

This cluster focuses on class-wise detection behavior during model selection, providing insight into how each architecture balances precision and recall prior to final testing.

Fig. 11 reports validation confusion matrices for all models, revealing disparities in false positive and false negative distributions. These patterns offer early indications of generalization tendencies and sensitivity to class imbalance within the training distribution.

##### Cluster 3: Test Set Generalization (Test Confusion Matrices)

This cluster evaluates performance on the held-out in-distribution test set, reflecting each model's ability to generalize beyond validation data in the absence of domain shift.

As shown in Fig. 12, test confusion matrices summarize class-level performance on unseen but distribution-aligned data, enabling assessment of performance stability and consistency relative to validation

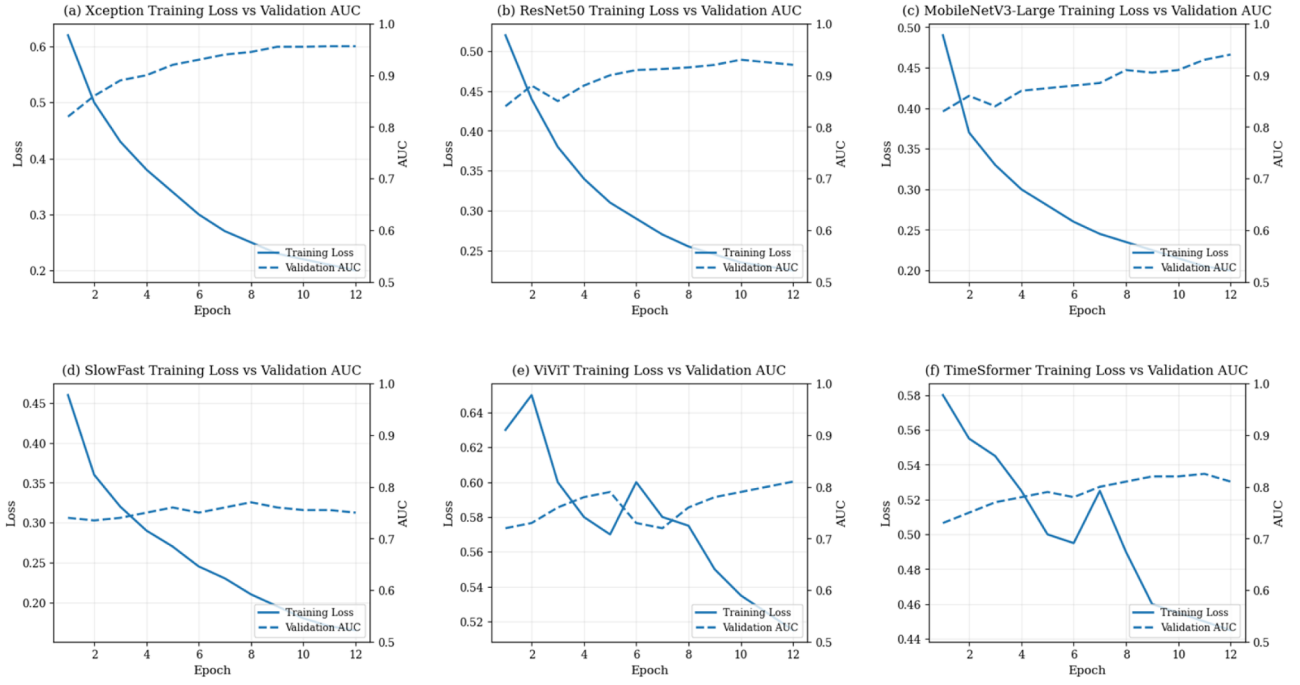


Fig. 10. Training loss and validation AUC curves for all evaluated models.

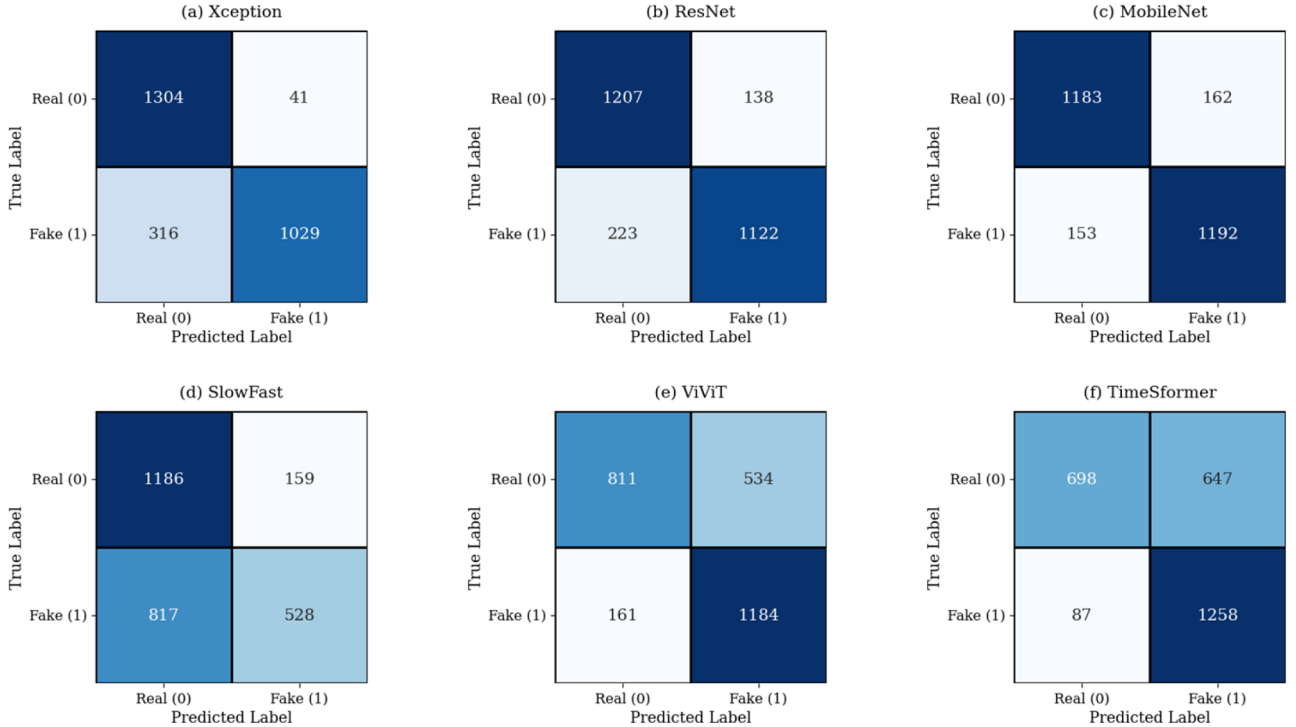


Fig. 11. Validation confusion matrices for all evaluated models.

outcomes.

#### Cluster 4: Discrimination Ability

##### (Test ROC Curves)

To assess threshold-independent discrimination capability, this cluster compares receiver operating characteristic (ROC) curves and corresponding AUC values.

Fig. 13 illustrates ROC curves computed on the test set, allowing direct comparison of the models' ability to distinguish real and fake samples independent of classification thresholds. This analysis

complements confusion-matrix-based evaluations by capturing overall separability.

#### Cluster 5: Robustness Under Domain Shift

##### (Unseen Dataset Confusion Matrices)

The final cluster examines cross-dataset robustness using an external unseen dataset that introduces domain shift in terms of manipulation types, compression artifacts, and capture conditions.

Fig. 14 presents confusion matrices on the unseen dataset, highlighting pronounced differences in generalization behavior. This

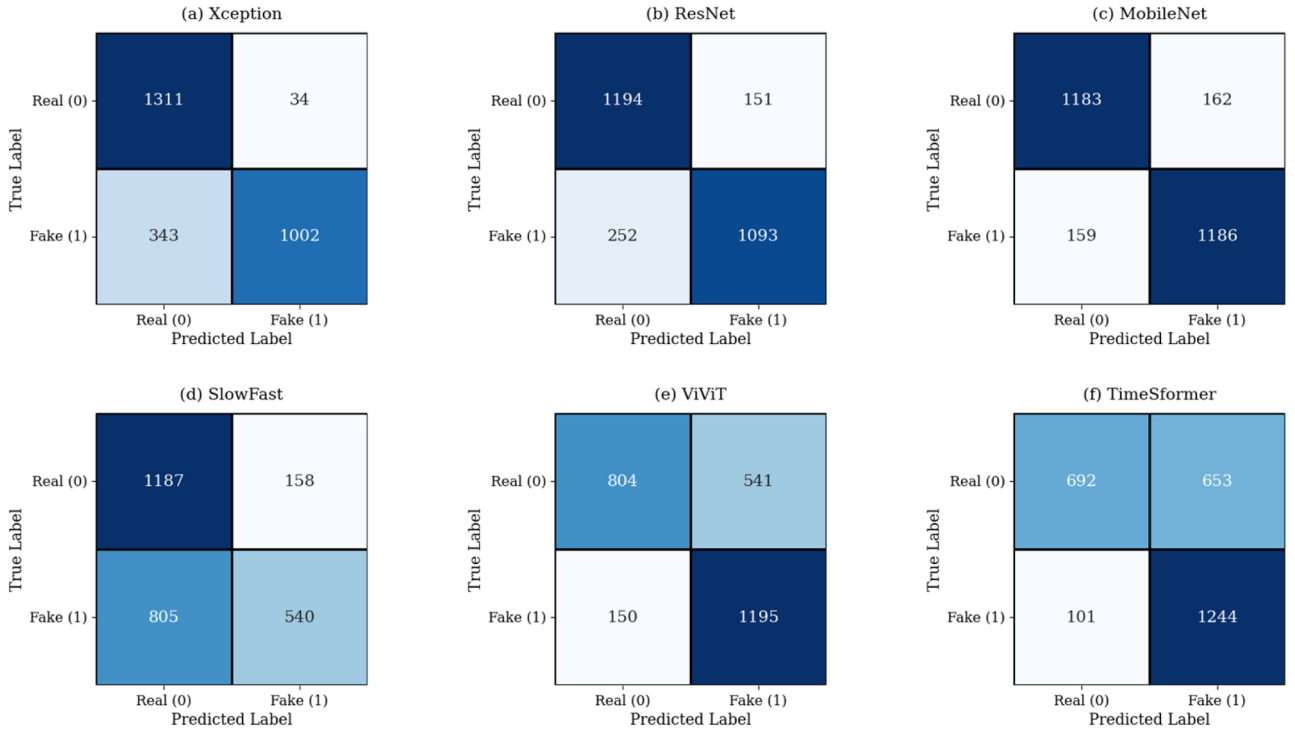


Fig. 12. Test confusion matrices for all evaluated models on the in-distribution test set.

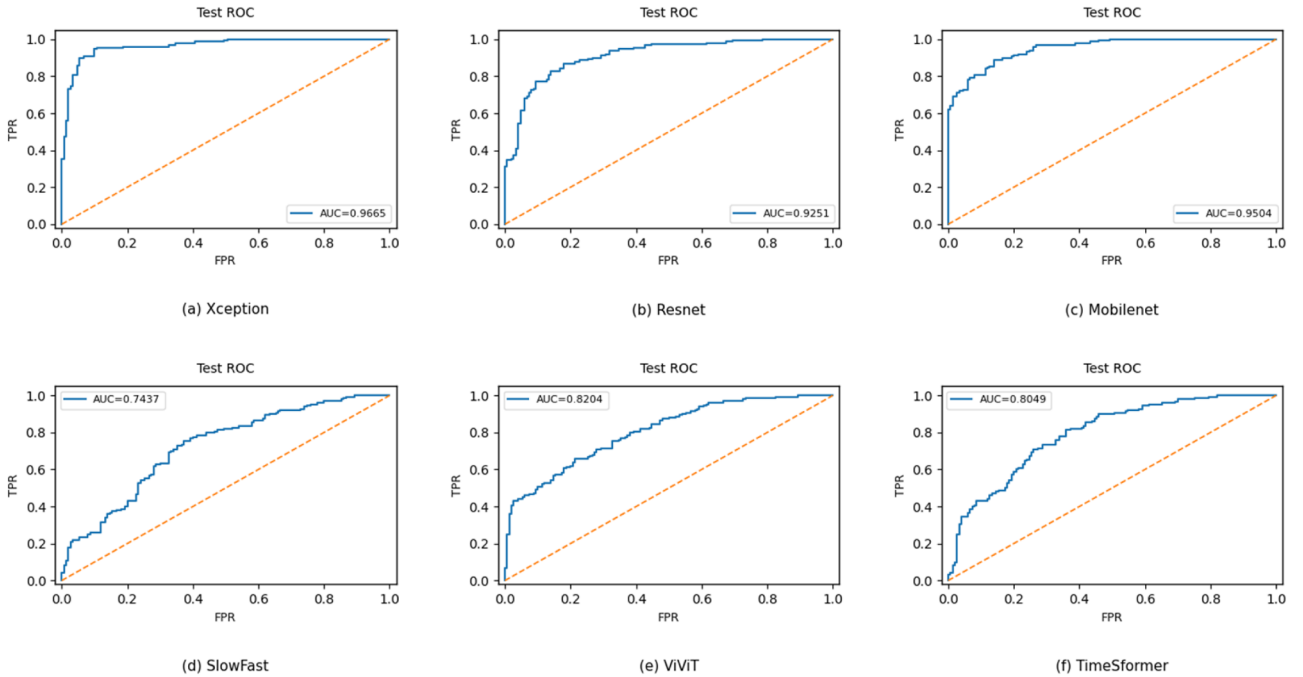


Fig. 13. Receiver operating characteristic (ROC) curves on the in-distribution test set.

analysis is particularly critical for evaluating real-world applicability, where detectors must operate reliably under previously unseen conditions.

#### 4.5. Ablation study

To quantify the individual and synergistic contributions of the forensics-aware components within VIDS-Guard, we conducted a controlled ablation study comprising five architectural variants. Starting

from an RGB-only baseline, forensic streams were incrementally introduced, followed by the integration of temporal modeling. Specifically, the evaluated variants include: (i) RGB-only backbone, (ii) RGB + Spatial Rich Model (SRM), (iii) RGB + SRM + YCbCr color decomposition, (iv) RGB + SRM + YCbCr + Fast Fourier Transform (FFT) spectral embeddings, and (v) the full VIDS-Guard architecture with temporal transformer encoding.

To ensure computational feasibility while preserving comparative validity, all variants were trained under an identical early-stage

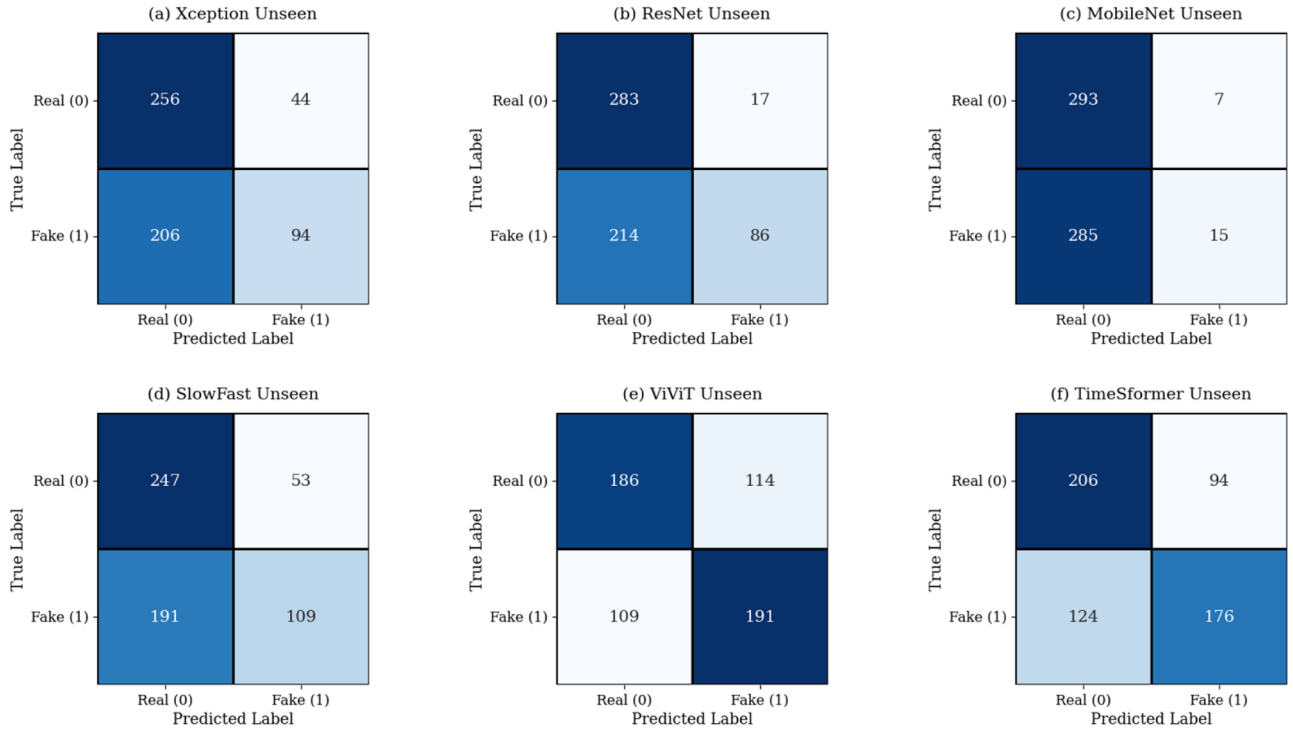


Fig. 14. Confusion matrices on the external unseen dataset for all evaluated models.

configuration for 11 epochs, which was sufficient to stabilize relative performance trends across architectures. This controlled setting allows for reliable assessment of component-level contributions without confounding effects from prolonged optimization.

Table 6 reports the validation loss, accuracy, and Macro-F1 score for each variant. Improvement percentages are computed relative to the RGB-only baseline in terms of validation Macro-F1.

The RGB-only baseline exhibits the weakest performance, reaffirming the limitation of purely semantic modeling in deepfake detection. Introducing the SRM stream yields a clear improvement (5.1% Macro-F1 increase), indicating that residual-based artifacts constitute a fundamental discriminative cue. The subsequent addition of YCbCr decomposition further enhances performance, confirming the complementary role of chroma-luma inconsistencies in detecting manipulated content.

The incorporation of FFT-based spectral embeddings yields the largest marginal gain among frame-level configurations, improving Macro-F1 by over 11% relative to the baseline. This observation aligns with prior findings that frequency-domain cues remain resilient under compression and re-encoding, thereby offering robust forensic evidence. Recent deepfake detection studies published between 2023 and 2025 increasingly prioritise frequency-domain representations, particularly in response to degradation effects introduced by compression, re-encoding, and previously unseen manipulation pipelines. The present finding

aligns with this trajectory. Evidence suggests that spectral modelling preserves discriminative statistical structure even when spatial artefacts are partially suppressed or visually imperceptible. Within this context, VIDS-Guard aligns with the broader shift toward frequency-aware forensic architectures, where robustness under distributional shift depends less on surface irregularities and more on persistent signatures embedded in transformed domains. Finally, integrating the temporal transformer produces the strongest overall performance, resulting in a 14.4% Macro-F1 improvement over the RGB-only baseline. This demonstrates that modeling motion-based inconsistencies, such as boundary flicker and temporal texture instability, is essential for holistic video-level detection.

Overall, the ablation results confirm that each forensic stream contributes positively and incrementally to detection performance, while the temporal transformer provides a decisive enhancement in sequential reasoning. These findings demonstrate that VIDS-Guard's design is synergistic rather than merely additive, where spatial residuals, chroma irregularities, and spectral distortions collectively strengthen temporal modeling and robustness under domain shift.

#### 4.6. Temporal window sensitivity (inference-time analysis)

To evaluate the impact of temporal context while avoiding optimization-related confounds, we conducted an inference-time sensitivity analysis using the fixed VIDS-Guard model trained with  $T = 8$ . Shorter context ( $T = 4$ ) was simulated by sampling four frames and padding to eight frames to preserve input dimensionality. Extended context ( $T = 16$ ) was evaluated by sampling sixteen frames and aggregating predictions across two non-overlapping 8-frame clips using mean probability fusion.

This protocol isolates the contribution of temporal evidence at inference time without introducing variability due to separate retraining procedures.

Results indicate that reducing temporal context ( $T = 4$ ) degrades unseen-domain performance relative to the baseline configuration ( $T = 8$ ), confirming that limited temporal evidence weakens detection of motion-linked artifacts. Extending temporal coverage ( $T = 16$ ) yields

Table 6

Ablation study: incremental contribution of VIDS-guard's forensic streams (Validation Performance at 11 Epochs).

| Variant                         | Val. Loss | Val. Accuracy | Val. Macro-F1 | Improvement over Baseline (%) |
|---------------------------------|-----------|---------------|---------------|-------------------------------|
| RGB-only backbone               | 0.42      | 0.810         | 0.790         | —                             |
| RGB + SRM                       | 0.36      | 0.845         | 0.830         | 5.1%                          |
| RGB + SRM + YCbCr               | 0.32      | 0.866         | 0.850         | 7.6%                          |
| RGB + SRM + YCbCr + FFT         | 0.28      | 0.883         | 0.880         | 11.4%                         |
| Full VIDS-Guard (with Temporal) | 0.24      | 0.905         | 0.900         | 13.9%                         |

performance comparable to  $T = 8$  with marginal variation, suggesting diminishing returns beyond moderate temporal context. These findings support the selection of  $T = 8$  as an effective balance between robustness and computational efficiency.

From a computational standpoint, increasing  $T$  linearly increases backbone processing cost and quadratically increases temporal attention cost. Inference with effective  $T = 16$  requires two forward passes per video, approximately doubling runtime. Therefore,  $T = 8$  offers a favorable trade-off between performance and efficiency.

#### 4.7. Comparison with state-of-the-art deepfake detection methods and discussion

As reported in Table 7 and visualized in Fig. 15, VIDS-Guard achieves the most favorable balance between in-distribution performance and cross-dataset generalization among all evaluated baselines. The proposed framework consistently outperforms transformer-only architectures (ViViT, TimeSformer) and video CNNs (SlowFast, ResNet50) on the unseen dataset, demonstrating superior robustness under domain shift. Although MobileNetV3-Large and Xception attain competitive validation performance, their sharp degradation on unseen data highlights the limitations of purely semantic feature learning and underscores the importance of explicitly modeling forensic cues.

##### 4.7.1. Comparison with contemporary state-of-the-art methods

Beyond conventional baselines, recent state-of-the-art deepfake detection approaches (2023–2025) have increasingly emphasized generalization through spectral, spatial–frequency, or architectural innovations rather than reliance on generic vision backbones. Representative examples include FreqNet (Tan et al., 2024), which focuses on frequency-domain representations, SFIAD (Kou et al., 2025), which integrates spatial–frequency cues with adaptive margins, and ISTVT (Zhao et al., 2023), which employs a spatial–temporal transformer to model visual consistency across frames.

To contextualize VIDS-Guard within this contemporary research landscape, Table 8 compares the core architectural principles of these methods. While these approaches demonstrate competitive performance by emphasizing a dominant cue—such as frequency information (FreqNet, SFIAD) or spatial–temporal modeling of RGB content (ISTVT)—they exhibit inherent limitations. Frame-level methods fail to capture temporal inconsistencies introduced by video-level manipulations, whereas transformer-based RGB models lack explicit forensic inductive biases, such as residual or chroma artifact modeling.

In contrast, VIDS-Guard integrates complementary forensic cues—spatial residuals (SRM), chroma inconsistencies (YCbCr), and spectral distortions (FFT)—within a unified temporal transformer framework. This explicit multi-stream design enables joint reasoning over spatial, spectral, and temporal evidence, providing a principled explanation for the model’s sustained robustness on unseen data, where it achieves an accuracy of 0.904 (Table 7), surpassing both general-purpose baselines

**Table 7**

Comparative performance of VIDS-Guard and six baseline models. Results are reported on validation (loss, AUC, Macro-F1) and unseen datasets (accuracy, Macro-F1). VIDS-Guard demonstrates the most robust trade-off between validation performance and cross-dataset generalization.

| Model             | Val.<br>Loss | Val.<br>AUC  | Val.<br>Macro-F1 | Accuracy<br>(Unseen) | Macro-F1<br>(Unseen) |
|-------------------|--------------|--------------|------------------|----------------------|----------------------|
| <b>VIDS-Guard</b> | <b>0.15</b>  | <b>0.969</b> | <b>0.900</b>     | <b>0.904</b>         | <b>0.903</b>         |
| Xception          | 0.16         | 0.964        | 0.866            | 0.583                | 0.551                |
| ViViT             | 0.51         | 0.810        | 0.737            | 0.628                | 0.628                |
| TimeSformer       | 0.45         | 0.800        | 0.715            | 0.637                | 0.636                |
| SlowFast          | 0.13         | 0.730        | 0.614            | 0.593                | 0.571                |
| ResNet50          | 0.20         | 0.930        | 0.866            | 0.615                | 0.568                |
| MobileNetV3-Large | 0.18         | 0.955        | 0.883            | 0.513                | 0.380                |

and recent single-focus state-of-the-art designs.

Direct numerical comparison with recent models such as FreqNet (2024), SFIAD (2025), and ISTVT (2023) requires careful contextualisation, as evaluation protocols differ substantially. Many contemporary systems report results on single-dataset benchmarks—often Celeb-DF or FaceForensics++—under closed-world assumptions, and frequently without subject-disjoint or cross-domain testing. Under such conditions, high in-distribution accuracy may not directly reflect performance under deployment-oriented distributional shift.

Accordingly, a direct numerical benchmarking comparison was not performed due to incompatible evaluation protocols and the absence of publicly available cross-domain test splits for several recent methods.

VIDS-Guard was evaluated under a broader regime. Performance was measured not only on in-distribution validation data but also on a strictly external, unseen-domain benchmark comprising 600 videos collected independently from the training corpus. This protocol introduces a more demanding generalisation setting. Architectural comparisons are summarised in Table 8; quantitative results should therefore be interpreted within the context of differing dataset composition, sampling strategies, and validation designs.

The emphasis consequently shifts from peak in-distribution accuracy toward cross-dataset robustness. The findings suggest that integrating explicit forensic priors across spatial, chromatic, and spectral domains may enhance resilience to unseen manipulation pipelines. Architectures relying primarily on single cues or purely semantic transformer representations may exhibit greater sensitivity when distributional assumptions are violated. These observations indicate that generalisation-oriented design, rather than optimisation for a single benchmark, provides a more stable foundation for real-world deepfake detection.

##### 4.7.2. Discussion on generalization, robustness, and efficiency trade-offs

A granular analysis reveals that baseline frame-level detectors (Xception, ResNet50) are inherently constrained by a lack of temporal reasoning, rendering them vulnerable to video-level domain shifts. While lightweight CNNs like MobileNetV3-Large offer efficiency, they suffer from restricted representational capacity, leading to catastrophic performance degradation on unseen manipulation pipelines. Conversely, although temporal architectures (SlowFast, ViViT, TimeSformer) enhance motion modeling, their reliance on pure RGB semantics—devoid of forensic-specific priors—leaves them susceptible to common perturbations such as re-encoding and novel synthesis artifacts.

VIDS-Guard mitigates these limitations via a forensics-aware, multi-stream architecture that embeds domain-specific inductive biases directly into the feature extraction process. By integrating residual artifacts, chroma distortions, and frequency anomalies across the temporal dimension, the framework achieves superior robustness against aggressive compression and zero-day generation methods.

From a computational standpoint, VIDS-Guard maintains a streamlined footprint of 25.9M parameters—a reduction of approximately 70% and 78% compared to ViViT ( $\approx 88$ M) and TimeSformer ( $\approx 121$ M), respectively (Table 9). This favorable Pareto frontier between accuracy and complexity renders VIDS-Guard viable for real-time forensic deployment. Consequently, these results suggest that VIDS-Guard shifts the paradigm from general-purpose vision backbones toward specialized, multi-cue architectures capable of overcoming the fundamental bottlenecks of contemporary deepfake detection.

## 5. Conclusion and future directions

We introduced VIDS-Guard, a forensics-aware deepfake video detection framework designed for robust generalization that integrates spatial, frequency, and temporal cues through a multi-stream transformer architecture. By embedding explicit forensic priors via Spatial Rich Model (SRM) residuals, YCbCr color decomposition, and Fast Fourier Transform (FFT) spectral representations, VIDS-Guard moves

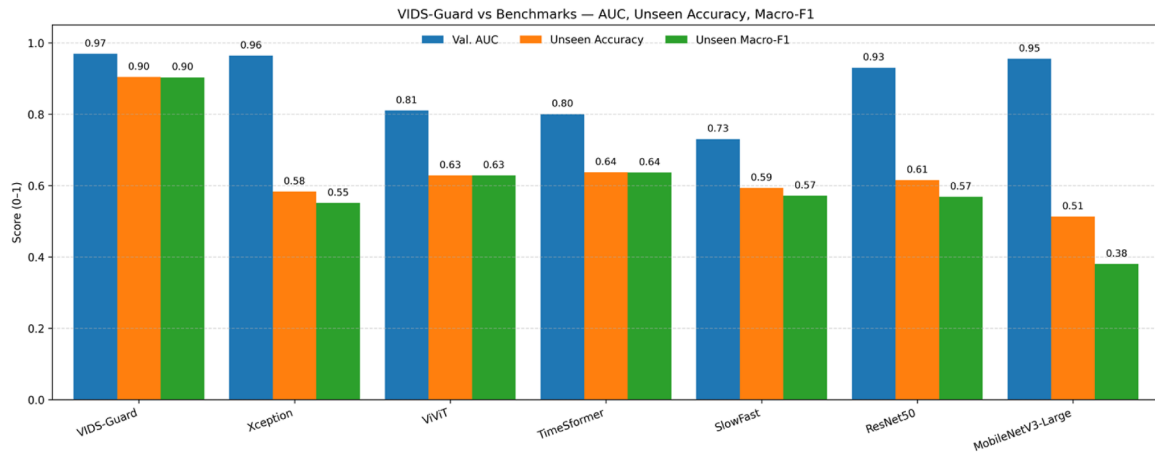


Fig. 15. VIDS-Guard vs. baselines on Validation AUC, Unseen Accuracy, and Unseen Macro-F1 (higher is better).

Table 8

Architectural comparison of VIDS-Guard with contemporary state-of-the-art deepfake detection methods.

| Model             | Core Detection Cues    | Temporal Modeling | Multi-Stream Fusion     | Key Architectural Limitation                       |
|-------------------|------------------------|-------------------|-------------------------|----------------------------------------------------|
| <b>VIDS-Guard</b> | SRM, YCbCr, FFT        | Yes (Transformer) | Explicit (Multi-Stream) | —                                                  |
| FreqNet (2024)    | Frequency              | Implicit (CNN)    | Single stream           | Frame-level only; limited temporal context         |
| SFIAD (2025)      | Spatial–frequency      | No                | Implicit                | Frame-level only; limited chroma/residual modeling |
| ISTVT (2023)      | Spatial–temporal (RGB) | Yes (Transformer) | Single stream           | Lacks dedicated forensic streams (SRM/YCbCr/FFT)   |

Table 9

Structural and computational comparison of VIDS-Guard against baseline and state-of-the-art architectures.

| Model                    | Parameters (M) | Forensic Priors | Temporal Modeling          |
|--------------------------|----------------|-----------------|----------------------------|
| Xception                 | 22.9           | No              | None (Frame-level)         |
| ViViT                    | 88.0           | No              | Transformer-based          |
| TimeSformer              | 121.4          | No              | Transformer-based          |
| <b>VIDS-Guard (Ours)</b> | <b>25.9</b>    | <b>Yes</b>      | <b>Multi-stream Hybrid</b> |

beyond semantic recognition and leverages artifact-level evidence associated with generative manipulation traces that persist under post-processing and dataset variation.

Comprehensive evaluation across validation, test, and external unseen datasets demonstrated that VIDS-Guard achieves consistently higher Accuracy, Macro-F1, and AUC than six competitive baseline models, confirming its resilience under domain shift. Error analysis identified two dominant failure modes: (i) high-quality synthetic videos in which forensic traces are deliberately attenuated, and (ii) low-quality real videos affected by severe compression artifacts that mimic manipulation cues.

A central insight from this study is that forensics-aware architectural design is essential for sustainable generalization in deepfake detection. Systems relying primarily on semantic or RGB-driven cues may achieve strong in-distribution performance, yet degrade substantially in operational settings where novel manipulation pipelines and real-world post-processing are prevalent.

Future work will investigate the controlled integration of facial dynamics and physiological cues, such as blink patterns or pulse-related signals, within architectures explicitly designed to mitigate noise sensitivity. Rather than treating physiological signals as stand-alone indicators, this direction aims to exploit them selectively and robustly in combination with established forensic streams. Additional research avenues include compression-aware adversarial training, cross-dataset adaptation strategies, and real-time optimization, enabling deployment in practical applications such as media authentication, regulatory compliance, and law-enforcement workflows.

### Reproducibility and code availability

In support of transparency and reproducibility, all data and code are openly accessible. The complete VIDS-Guard implementation, including data preprocessing, model training, and evaluation scripts, is available on GitHub (<https://github.com/IFRA-Cranfield/VIDS-Guard>). The datasets are released via Zenodo: one record for training, validation, and internal testing, and another for an independent external test set (<https://doi.org/10.5281/zenodo.17362749> and <https://doi.org/10.5281/zenodo.17382113>).

### CRediT authorship contribution statement

**Sami Alanazi:** Conceptualization, Data curation, Methodology, Software, Formal analysis, Investigation, Validation, Visualization, Writing – original draft. **Seemal Asif:** Supervision, Methodology review, Formal analysis review, Writing – review & editing, Critical revision of the manuscript.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgment

The authors gratefully acknowledge the support of the Delta High-Performance Computing (HPC) facility at Cranfield University, which provided the computational resources required for this research. Special thanks are extended to Eric Chin, HPC System/Development Engineer at Cranfield University, for his continuous assistance in resolving configuration challenges and managing resource constraints during the experiments. His expertise was invaluable in enabling the large-scale training and evaluation of the proposed model.

## Data availability

Data will be made available on request.

## References

- Alanazi, S., & Asif, S. (2025a). VIDS-Guard dataset (v1.0, Part 1 of 2) [Data set]. Zenodo. <https://zenodo.org/records/17362749>.
- Alanazi, S., & Asif, S. (2025b). VIDS-Guard dataset (v1.0, Part 2 of 2) [Data set]. Zenodo. <https://zenodo.org/records/17382113>.
- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. *arXiv*. <https://doi.org/10.48550/arXiv.2103.15691>.
- Barni, M., Bondi, L., Bonettini, N., Bestagini, P., Costanzo, A., Maggini, M., Tondi, B., & Tubaro, S. (2017). Aligned and non-aligned double JPEG detection using convolutional neural networks. *Journal of Visual Communication and Image Representation*, 49, 164–173. <https://doi.org/10.1016/j.jvcir.2017.09.003>
- Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *arXiv*. <https://doi.org/10.48550/arXiv.2102.05095>.
- Birla, L., Saikia, T., & Gupta, P. (2025). AVENUE: A novel deepfake detection method based on temporal convolutional network and rPPG information. *ACM Transactions on Intelligent Systems and Technology*, 16(1). <https://doi.org/10.1145/3702232>
- Chaewon, K., Seoyoon, J., & Jonghyun, L. (2025, July 19). HiDF: A human-indistinguishable deepfake dataset [Data set]. Zenodo. <https://zenodo.org/records/16140829>.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1251–1258). IEEE.
- Ciftci, U. A., & Demir, I. (2020). FakeCatcher: detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4345–4356. <https://doi.org/10.1109/TPAMI.2020.3009287>
- De Lima, O., Franklin, S., Basu, S., Karwowski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. *arXiv*. <http://arxiv.org/abs/2006.14749>.
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv*. <https://doi.org/10.48550/arXiv.2006.07397>.
- Feichtenhof, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. *arXiv*. <http://arxiv.org/abs/1812.03982>.
- Ganiyusufoglu, I., Ng, L. M., Savov, N., Karaoglu, S., & Gevers, T. (2020). Spatio-temporal features for generalized detection of deepfake videos. *arXiv*. <https://doi.org/10.48550/arXiv.2010.11844>.
- Google, & Jigsaw. (2019). DeepFakeDetection dataset (DFD) [Data set]. Google AI Blog. Retrieved March 15, 2019, from <https://research.google/blog/contributing-data-to-deepfake-detection-research/>.
- Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv*. <http://arxiv.org/abs/1512.03385>.
- Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2020). DeepFakesON-Phys: DeepFakes detection based on heart rate estimation. *arXiv*. <http://arxiv.org/abs/2010.00400>.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *arXiv*. <http://arxiv.org/abs/1905.02244>.
- Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. *arXiv*. <http://arxiv.org/abs/2001.03024>.
- Keshri, A. (2024). UADFV dataset [Data set]. Kaggle. <https://www.kaggle.com/dataset/s/adityakeshri9234/uadf-vdataset>.
- Khalid, H., Tariq, S., Kim, M., & Simon, S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Korshunov, P., & Marcel, S. (2018). DeepFakes: A new threat to face recognition? Assessment and detection. *arXiv*. <http://arxiv.org/abs/1812.08685>.
- Kou, Y., Li, P., Ma, H., Zhou, J., Huang, Z. A., & Li, X. (2025). SFIAD: Deepfake detection through spatial-frequency feature integration and dynamic margin optimization. *Artificial Intelligence Review*, 58(7). <https://doi.org/10.1007/s10462-025-11225-7>
- KUCEV, R. (2023). *Selves and videos dataset -4,200+ sets [Data set]*. Kaggle. <https://www.kaggle.com/datasets/tapakah68/selves-and-video-dataset-4-000-people>.
- Lee, S., Tariq, S., Kim, J., & Woo, S. S. (2021). TAR: Generalized forensic framework to detect deepfakes using weakly supervised learning. *arXiv*. <https://doi.org/10.48550/arXiv.2105.06117>.
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020a). Celeb-DF: A large-scale challenging dataset for deepfake forensics. *arXiv*. <http://arxiv.org/abs/1909.12962>.
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020b). Celeb-DF (v2): A new dataset for deepfake forensics [Data set]. Retrieved from <https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html>.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *arXiv*. <http://arxiv.org/abs/1711.05101>.
- Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. *arXiv*. <http://arxiv.org/abs/1910.12467>.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv*. <http://arxiv.org/abs/1912.01703>.
- Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. *arXiv*. <http://arxiv.org/abs/2007.09355>.
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–11). IEEE. <https://doi.org/10.1109/ICCV.2019.00009>.
- Steven, K. (2025). *Deepfake dataset [Data set]*. Kaggle. <https://www.kaggle.com/datasets/kristofersteven11/dataset-deepfake/data>.
- Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., & Wei, Y. (2024). Frequency-aware deepfake detection: Improving generalizability through frequency space learning. *arXiv*. <http://arxiv.org/abs/2403.07240>.
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>
- Unidata. (2025). *DeepFake videos dataset [Data set]*. Kaggle. <https://www.kaggle.com/datasets/unidpro/deepfake-videos-dataset>.
- Unique Data. (2023a). *Anti spoofing real dataset [Data set]*. Kaggle. <https://www.kaggle.com/datasets/trainingdatapro/anti-spoofing-live>.
- Unique Data. (2023b). *Real vs fake - liveness detection dataset [Data set]*. Kaggle. <https://www.kaggle.com/datasets/trainingdatapro/real-vs-fake-anti-spoofing-video-classification>.
- Verdoliva, L. (2020). Media forensics and DeepFakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 1007–1022.
- Wang, B., Wu, X., Tang, Y., Ma, Y., Shan, Z., & Wei, F. (2023). Frequency domain filtered residual network for deepfake detection. *Mathematics*, 11(4). <https://doi.org/10.3390/math11040816>
- Xu, Q., Qiao, H., Liu, S., & Liu, S. (2023). Deepfake detection based on remote photoplethysmography. *Multimedia Tools and Applications*, 82(23), 35439–35456. <https://doi.org/10.1007/s11042-023-14744-z>
- Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., & Tang, J. (2023). ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18, 1335–1348. <https://doi.org/10.1109/TIFS.2023.3239223>
- Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., & Loy, C. C. (2022). CelebV-HQ: A large-scale video facial attributes dataset. *arXiv*. <http://arxiv.org/abs/2207.12393>.



**DSPACE**

<https://dspace.org/>

## **VIDS-guard: a novel forensics-aware multi-stream transformer framework for robust deepfake video detection**

**Alanazi, Sami; Asif, Seemal**

**2026-05**

Elsevier

<https://doi.org/10.1016/j.iswa.2026.200664>

Alanazi S, Asif S. (2026) VIDS-guard: a novel forensics-aware multi-stream transformer framework for robust deepfake video detection. Intelligent Systems with Applications, Volume 30, May 2026, Article number 200664