International Journal of Intelligent Systems

WILEY

RESEARCH ARTICLE OPEN ACCESS

Beyond Frame-Level Accuracy: Analyzing Video-Level Stability in Transformer-Based Deepfake Forensics

Syed Azmul Hasan[1] | Mowmita Pervin Hera[1] | Afroza Lona Mim[1] | M. M. Jihad Hossain[1] | Hazrat Ali[2] | Mohammad Farhad Bulbul[1]

[1]Department of Mathematics, Jashore University of Science and Technology, Jashore 7408, Bangladesh | [2]School of Computing, Data and Mathematical Sciences, University of Stirling, Stirling, UK

Correspondence: Mohammad Farhad Bulbul (farhad@just.edu.bd)

Received: 3 April 2026 | **Revised:** 27 July 2026 | **Accepted:** 29 September 2026

Keywords: Celeb-DF v2 | convolutional neural networks | deepfake detection | FaceForensics++ | Vision Transformer

ABSTRACT

Deepfake videos generated by advanced generative models pose serious threats to digital security, misinformation control, and media forensics. Although numerous detection methods have been proposed, their reported performances are often based on heterogeneous preprocessing pipelines, inconsistent data splits, and model-specific training protocols, hindering fair architectural comparison. This work presents a controlled and reproducible comparative study of convolutional neural networks (ResNet-18 and EfficientNet-B0) and a Transformer-based architecture (ViT-B/16) within a unified, face-centric deepfake detection framework. All models are trained using identical preprocessing, augmentation, optimization settings, and data splits on Celeb-DF v2 and FaceForensics++, and results are validated across three independent random seeds with bootstrapped confidence intervals. Experimental results show that while CNN-based models perform competitively at the frame level, their performance degrades after video-level aggregation. In contrast, ViT-B/16 demonstrates more stable video-level performance, achieving 92.8% (95% CI: 91.7–94.0) video-level accuracy and an area under the curve (AUC) of 0.950 (95% CI: 0.934–0.964) on Celeb-DF v2 (3-seed bootstrapped). Cross-dataset evaluation across the same three random seeds reveals a large, asymmetric generalization gap (AUC 0.79 FF++ $\longrightarrow$ Celeb-DF vs. 0.64 Celeb-DF $\longrightarrow$ FF++), indicating that in-domain video-level stability does not transfer to unseen manipulation domains. These findings indicate that global self-attention improves robustness and video-level stability for deepfake detection within a given data domain, while highlighting the need for further work on cross-domain generalization.

1 | Introduction

Recent advances in generative modeling—particularly generative adversarial networks (GANs), autoencoder-based pipelines, and neural face-editing frameworks—have enabled the synthesis of highly realistic facial images and videos, commonly referred to as deepfakes. While such technologies have legitimate applications in entertainment and media production, their misuse poses serious risks to digital security, misinformation control, and personal privacy [1, 2]. Continuous improvements in visual realism and temporal coherence have made modern deepfakes increasingly difficult to distinguish from authentic media.

Deepfake detection has, therefore, emerged as an important research area at the intersection of computer vision and multimedia forensics. Early detection approaches relied on handcrafted artifacts such as eye-blinking inconsistencies and head-pose irregularities [3, 4]. However, as generative models improved, these low-level cues became less reliable. Consequently,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Copyright © 2026 Syed Azmul Hasan et al. *International Journal of Intelligent Systems* published by John Wiley & Sons Ltd.

research shifted toward deep learning–based detectors trained on large-scale manipulated datasets. Benchmarks such as FaceForensics++ and Celeb-DF v2 have enabled standardized evaluation and accelerated progress in this field [1, 5].

Despite substantial advances, several challenges remain. Many detectors exhibit limited cross-dataset generalization and performance degradation under compression or unseen manipulation methods [6]. Furthermore, most existing approaches rely on convolutional neural networks (CNNs), including EfficientNet and ResNet variants [7, 8]. While CNNs effectively capture local texture artifacts, their limited receptive fields restrict their ability to model long-range spatial dependencies. Prior work suggests that incorporating global context can improve robustness to subtle and spatially distributed manipulations [9, 10].

Vision Transformers (ViTs) have demonstrated competitive or superior performance compared to CNNs across various visual recognition tasks [11]. In deepfake detection, global self-attention may enhance robustness by integrating evidence across multiple facial regions rather than focusing solely on localized texture inconsistencies. However, many Transformer-based detectors are evaluated under different preprocessing pipelines and training protocols, making a fair architectural comparison difficult.

Motivated by these limitations, this paper presents a controlled backbone-level comparison between CNN architectures (EfficientNet-B0 and ResNet-18) and a Transformer-based model (ViT-B/16) for deepfake detection. All models are trained within a unified face-centric pipeline using identical preprocessing, augmentation, data splits, and optimization settings on Celeb-DF v2 [5] and FaceForensics++ [1]. By eliminating experimental inconsistencies, we isolate architectural differences and provide a fair comparative evaluation.

Beyond frame-level analysis, we investigate the impact of video-level aggregation by averaging frame-wise prediction scores to obtain stable video-level decisions. We further analyze model interpretability using Grad-CAM for CNNs and attention visualization for transformers to examine spatial activation patterns.

The main contributions of this work are as follows:

- We present a controlled and reproducible backbone-level comparison of CNN and Transformer architectures under identical experimental conditions.
- We provide a systematic analysis of frame-level versus video-level performance and aggregation stability.
- We provide a comparative interpretability study highlighting differences in spatial activation patterns.
- We provide empirical insights into the robustness and computational trade-offs for practical deployment.
- We provide a 3-seed, statistically validated cross-dataset generalization study, quantifying the extent to which video-level stability transfers across domains.
- We empirically test an explicit temporal-modeling alternative (a Gated Recurrent Unit [GRU]-based aggregation) against mean pooling, reporting a preliminary negative result that motivates directions for future work.

The remainder of this paper is organized as follows. Section 2 reviews prior work on deepfake generation and detection, covering CNN-based, frequency-domain, temporal, and Transformer-based approaches. Section 3 presents the proposed unified face-centric detection framework, detailing preprocessing, backbone architectures, and training protocol. Section 4 reports quantitative and qualitative experimental results, including frame-level and video-level evaluations, statistical robustness, and cross-dataset generalization. Section 5 discusses the impact of global self-attention, aggregation behavior, interpretability, computational trade-offs, and limitations. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2 | Related Work

The rapid advancement of generative models, such as GANs and neural rendering techniques, has enabled the creation of highly realistic manipulated videos, commonly referred to as deepfakes. While these technologies offer promising applications in entertainment and media production, they introduce severe risks regarding misinformation and digital security. Consequently, extensive research has shifted toward developing robust, automated detection methods leveraging computer vision and deep learning.

2.1 | Artifact-Based Spatial Detection

Early detection methods primarily focused on identifying visual artifacts introduced during the synthesis process, such as inconsistencies in blending, boundary generation, and texture synthesis. The field saw substantial progress with the introduction of large-scale benchmarks, particularly FaceForensics++ by Rössler et al. [1]. This study demonstrated that deep CNNs, specifically XceptionNet, could achieve high accuracy by treating deepfake detection as a per-frame classification task.

Building on this, researchers explored more sophisticated architectures to enhance feature representation. Lacerda et al. [12] utilized EfficientNet-based models to improve performance through compound scaling and parameter efficiency. Meanwhile, Wang et al. [13] conducted a comprehensive study on CNN-based detectors, highlighting the significant challenges of generalization across different generative models. Alternative architectures like Capsule-Forensics [14] were also proposed to preserve hierarchical relationships between facial features. To target specific synthesis flaws, Li et al. [15] introduced Face X-ray, which explicitly models blending boundaries to identify manipulation without prior knowledge of the synthesis method.

2.2 | Spatiotemporal and Physiological Approaches

While spatial methods are effective at capturing frame-level artifacts, they often overlook temporal inconsistencies. To address this, researchers integrated temporal dynamics and physiological cues. Qi et al. [16] introduced DeepRhythm, which utilizes remote photoplethysmography (rPPG) to detect subtle, heartbeat-induced variations in facial regions—a biological signal that generative models struggle to replicate. Similarly, Prashnani et al. [17] proposed PhaseForensics, leveraging phase-based motion representations and 3D CNNs to capture temporal irregularities in motion that are imperceptible in the spatial domain.

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Furthermore, lightweight alternatives have been explored for computational efficiency. Chen et al. [18] developed DefakeHop, a framework based on successive subspace learning. Unlike standard deep networks, DefakeHop extracts hierarchical features through a data-driven approach without backpropagation, maintaining competitive performance on benchmarks like Celeb-DF [5] while requiring significantly fewer parameters.

2.3 | Attention and Multidomain Feature Learning

To improve the localization of subtle artifacts, recent studies have incorporated attention mechanisms. Zhao et al. [19] proposed a multiattentional framework using multiple spatial attention heads to capture artifacts across various facial regions simultaneously. Similarly, Omar et al. [20] utilized self-attention mechanisms to dynamically emphasize discriminative spatial patterns, improving accuracy in high-quality, low-artifact scenarios.

Beyond attention, multidomain learning has gained traction. Xu et al. [21] introduced MCX-API, which integrates multichannel color-space representations with pairwise interaction learning. By capturing cross-channel dependencies, this approach enhances the model's ability to generalize across diverse datasets, bridging the performance gap between FaceForensics++ and Celeb-DF.

2.4 | Benchmark Datasets and Evaluation

The evolution of detection algorithms has been driven by benchmark datasets. FaceForensics++ [1] remains the standard for controlled manipulation studies, while Celeb-DF v2 [5] provides a more challenging and high-quality benchmark with fewer visual artifacts. Due to the high realism of these datasets, performance is standardly evaluated using the area under the curve (AUC)–receiver operating characteristic (ROC) to ensure a threshold-independent assessment.

2.5 | Summary and Research Gap

Despite the progress in CNN-based and spatiotemporal methods, two primary challenges remain. Most CNNs overfit to specific dataset artifacts and fail when tested on unseen manipulation techniques. In addition, traditional architectures often focus on local textures and fail to capture global structural dependencies across the entire face.

To address these limitations, recent research has shifted toward Transformer-based architectures, which utilize self-attention to capture long-range dependencies and global contextual relationships. Motivated by these advantages, this work proposes a unified framework built on the ViT-B/16 architecture. By employing a consistent preprocessing pipeline and leveraging the global receptive field of Transformers, the proposed method aims to achieve superior robustness and cross-dataset generalization on both the Celeb-DF and FaceForensics++ benchmarks.

3 | Methodology

The proposed pipeline consists of face-centric preprocessing, frame-level classification, and video-level aggregation under identical experimental settings. In addition, this section describes the cross-dataset evaluation protocol (§3.5), the temporal aggregation ablation (§3.6), and the statistical evaluation procedure (§3.7) used to validate all reported results.

3.1 | System Overview

Given an input video, frames are uniformly sampled at fixed temporal intervals to ensure consistent coverage while maintaining computational efficiency. Each sampled frame is processed using MTCNN for face detection, and the largest detected face is cropped and resized to 224×224. The cropped region of the face is passed to a backbone network (ResNet-18, EfficientNet-B0, or ViT-B/16), which outputs a frame-level manipulation probability p_i. For a video containing K sampled frames, predictions are aggregated using mean pooling as follows:

$$P_{\text{video}} = \frac{1}{K}\sum_{k=1}^{K} p_k, \tag{1}$$

where P_{video} denotes the final video-level prediction. A threshold $\tau = 0.5$ is used for binary classification. Figure 1 shows the overall workflow of the pipeline.

3.2 | Datasets and Preprocessing

Experiments were conducted on Celeb-DF v2 [5] and FaceForensics++ [1], two widely adopted deepfake benchmarks. Both datasets are split at the video level into 80% training, 10% validation, and 10% testing sets to prevent data leakage. Exact split sizes and per-split real/fake frame counts are reported in Table 1.

Frames are uniformly sampled from each video. Detected faces are resized to 224×224 and normalized using ImageNet statistics. Data augmentation includes random resized cropping, horizontal flipping, color jitter, and optional Gaussian blur or noise to improve robustness against compression artifacts. We note that both datasets are class-imbalanced and skewed toward the fake class (Celeb-DF v2: 84%–87% fake frames and FaceForensics++: 65%–69% fake frames), and accuracy figures reported in Section 4 should, therefore, be interpreted alongside F1 and AUC rather than in isolation.

3.3 | Model Architectures

Three backbone architectures were evaluated under identical preprocessing and optimization settings.

CNN-based models: ResNet-18 and EfficientNet-B0 were initialized with ImageNet-pretrained weights and fine-tuned for binary classification.

Transformer-based model: ViT-B/16 partitions each input image into 16×16 patches and processes them using 12 transformer encoder layers with multihead self-attention. The final classification token is used to predict manipulation probability. Figure 2 shows the architecture summary of the ViT model used in this work.

3.4 | Training Protocol

All models are trained using binary cross-entropy loss and the AdamW optimizer with layer-wise learning-rate decay of 0.7 and a base learning rate of 1.4×10^{-5}, weight decay of 0.06, and an effective batch size of 32 (microbatch size 16 with gradient

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

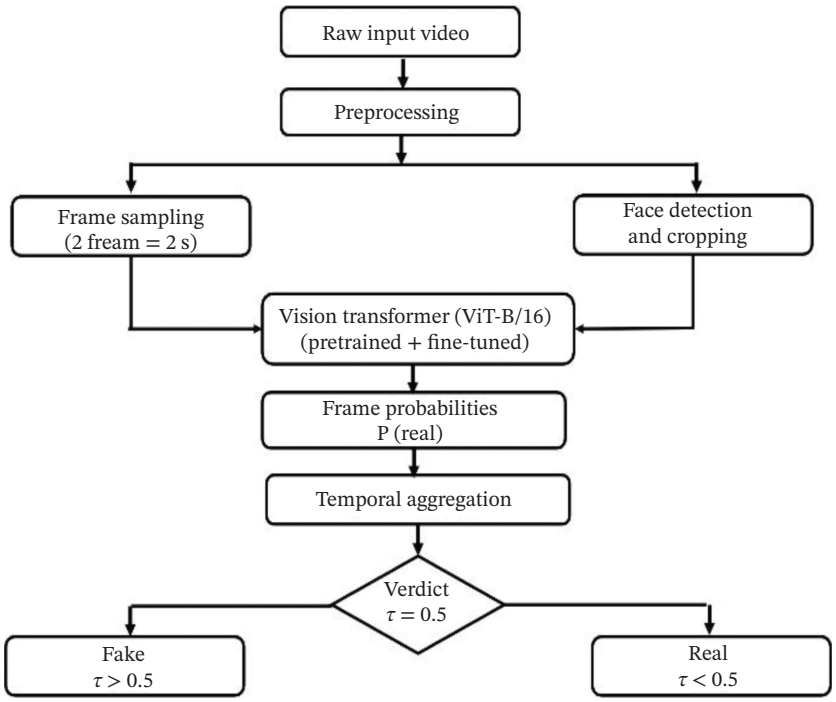

FIGURE 1 | Overall workflow of the proposed deepfake detection pipeline.

TABLE 1 | Dataset split statistics, including class balance.

Dataset	Split	Videos	Real frames	Fake frames	Total frames	Fake ratio (%)
FF++	Train	5600	71,545	136,784	208,329	65.7
FF++	Val	700	9078	17,065	26,143	65.3
FF++	Test	700	8157	18,164	26,321	69.0
Celeb-DF v2	Train	5223	6901	45,320	52,221	86.8
Celeb-DF v2	Val	653	1040	5490	6530	84.1
Celeb-DF v2	Test	653	950	5580	6530	85.5

accumulation of two). Models are trained for 10 epochs on FaceForensics++ and 6 epochs on Celeb-DF v2, with the checkpoint selected by best exponential-moving-average (EMA) validation ROC–AUC. All experiments were run on a single Tesla T4 GPU (14.56 GB VRAM) using PyTorch 2.10.0 + cu128 and timm 1.0.26.

To ensure fair comparison, identical data splits, augmentation strategies, and evaluation metrics were used across all architectures. To assess reproducibility and seed sensitivity, every configuration (both datasets, both cross-dataset directions, and the mean-pooling backbone) was trained and evaluated independently under three random seeds (42, 123, and 2024); results are aggregated and reported with confidence intervals in Section 4, as detailed in §3.7. Full hyperparameter and hardware settings are summarized in Table 2 for reproducibility.

3.5 | Cross-Dataset Evaluation Protocol

To evaluate generalization beyond the training domain, we perform cross-dataset testing by evaluating each trained model zero-shot on the held-out test split of the other dataset, without any fine-tuning or adaptation. Specifically, models trained on the full FaceForensics++ training split are evaluated on the Celeb-DF v2 test split (FF++ ⟶ Celeb-DF) and vice versa (Celeb-DF ⟶ FF++). The same face-centric preprocessing pipeline (MTCNN detection, 224 × 224 cropping, ImageNet normalization) is applied in both directions, with no dataset-specific tuning. For evaluation compute efficiency, a frame cap of 15 frames per video is used during cross-dataset inference. This protocol is repeated independently for all three random seeds in both directions, yielding six cross-dataset evaluation runs in total. Results are reported in §4.1.

3.6 | Temporal Aggregation Ablation

Video-level aggregation in the main pipeline (Equation (1)) uses simple mean pooling over frame-level probabilities, which does not explicitly model temporal dependencies between frames. To test whether explicit temporal modeling improves upon this simple aggregation strategy, we implement a GRU-based [22] temporal head as an alternative aggregator. The GRU operates on frozen frame-level embeddings extracted from the ViT-B/16 backbone, using a sequence length of 16 frames per video,

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

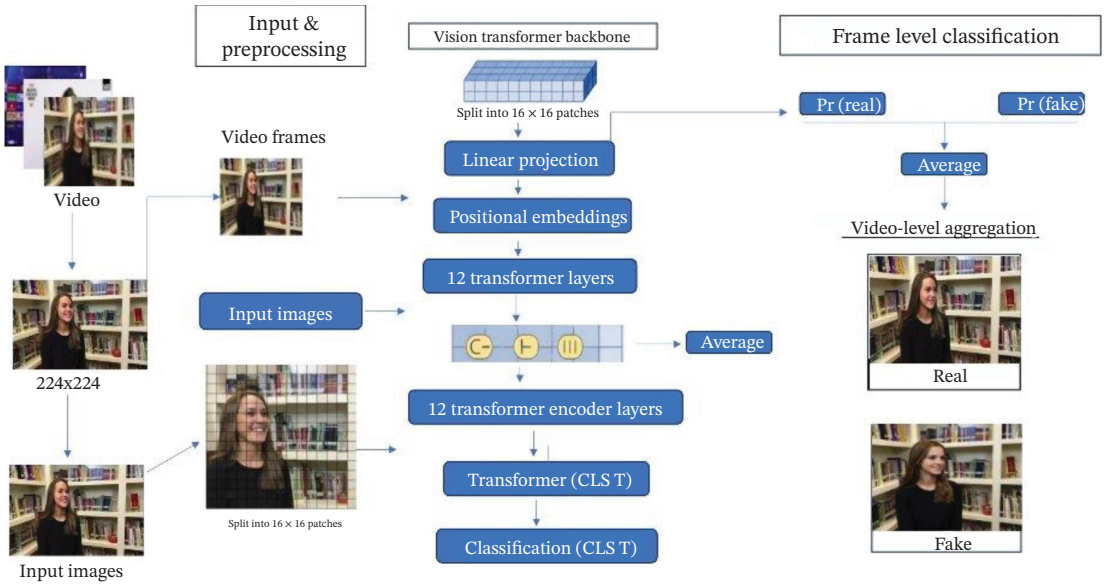

FIGURE 2 | Architecture of ViT-B/16 used for deepfake detection.

TABLE 2 | Reproducibility and hyperparameter settings.

Item	Value
Hardware	1 × Tesla T4 GPU, 14.56 GB VRAM
Software	PyTorch 2.10.0 + cu128, timm 1.0.26
Random seeds	42, 123, 2024 (train + in-domain evaluation + both cross-dataset directions)
Data split	80%/10%/10% (train/validation/test), split at video level
Image resolution	224 × 224
Batch size	16 (microbatch) × 2 (gradient accumulation) = 32 effective
Epochs (mean pool)	10 (FF++), 6 (Celeb-DF v2)
Epochs (temporal GRU)	8 (FF++ only, single seed, exploratory)
Optimizer	AdamW, layer-wise learning-rate decay = 0.7
Base learning rate	1.4×10^{-5} (backbone); 1×10^{-4} (temporal GRU head)
Weight decay	0.06
Checkpoint selection	Best EMA validation AUC
Temporal head	GRU, hidden size 256, sequence length 16 frames
Statistical testing	Bootstrap resampling (2000 iterations), 95% confidence interval, pooled across three seeds

a hidden size of 256, and is trained with the AdamW optimizer at a learning rate of 1×10^{-4} for 8 epochs. This ablation was run on FaceForensics++ using a single random seed (42) due to time constraints and is reported as a preliminary, exploratory result rather than a fully validated finding; it was not repeated across all three seeds or on Celeb-DF v2. Results are reported in §4.1 and discussed in §5.2.

3.7 | Statistical Evaluation

All in-domain and cross-dataset results are additionally evaluated using bootstrap resampling. For each experimental condition (in-domain FF++, in-domain Celeb-DF v2, and both cross-dataset directions), video-level predictions from all three completed random seeds are pooled, and 2000 bootstrap iterations are performed by resampling videos with replacement to compute 95% confidence intervals for accuracy, F1 score, and ROC–AUC. This procedure allows us to report point estimates alongside confidence intervals (Table 3 for in-domain results and Table 4 for cross-dataset results), demonstrating statistical significance analysis and seed-stability evidence.

3.8 | Evaluation Metrics

Performance is evaluated at both frame and video levels using accuracy, F1 score, and ROC–AUC. Video-level predictions are obtained via mean aggregation of frame-level probabilities, except where explicitly noted otherwise (§3.6). See §3.7 for the statistical evaluation methodology underlying the confidence intervals reported alongside these metrics in Section 4.

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | In-domain results, bootstrapped (2000 iterations, 95% CI, pooled across all 3 seeds).

Dataset	*N* videos	Acc (mean [95% CI])	F1 (mean [95% CI])	AUC (mean [95% CI])
FF++	2100	0.7741 [0.7557, 0.7910]	0.8353 [0.8207, 0.8489]	0.8404 [0.8206, 0.8586]
Celeb-DF v2	1959	0.9283 [0.9168, 0.9398]	0.9578 [0.9507, 0.9647]	0.9499 [0.9340, 0.9643]

TABLE 4 | Cross-dataset generalization, bootstrapped (2000 iterations, 95% CI, pooled across all 3 seeds).

Direction	*N* videos	Accuracy	F1	AUC
FF++ ⟶ Celeb-DF v2	1959	0.6691 [0.6483, 0.6901]	0.7685 [0.7505, 0.7861]	0.7881 [0.7617, 0.8133]
Celeb-DF v2 ⟶ FF++	2100	0.3877 [0.3671, 0.4100]	0.3133 [0.2851, 0.3407]	0.6386 [0.6141, 0.6642]

4 | Experimental Results

4.1 | Quantitative Evaluation

First, we evaluated the comparative performance of CNN-based baselines and the proposed ViT-B/16. Table 5 summarizes the configuration of the ViT-B/16 model used in all experiments.

To demonstrate statistical robustness, all ViT-B/16 results in the following were obtained across three independent random seeds (42, 123, and 2024) and are reported as bootstrapped estimates (2000 resampling iterations, 95% CI, pooled across seeds; see §3.7). CNN baselines (ResNet-18 and EfficientNet-B0) were trained once under the same data splits and preprocessing and are reported as single-run results; the primary purpose of the multiseed protocol is to validate the stability of the proposed Transformer-based pipeline, not to rebenchmark the baselines. Video-level performance results are presented in Table 6. Frame-level computational performance is reported in Table 7, including peak GPU memory usage measured separately during a single forward (inference) pass and during a full forward–backward (training) step, both including face-detection preprocessing overhead.

4.1.1 | Statistical Robustness

To provide statistical validation of the reported results (§3.7), Table 8 reports the raw per-seed accuracy, F1, and AUC values obtained independently under each of the three random seeds; these three per-seed runs are the underlying data that were pooled and bootstrapped to produce the ViT-B/16 summary rows in Table 6. Table 3 reports the resulting bootstrapped point estimates and 95% confidence intervals in full. All three seeds completed the full in-domain training and evaluation pipeline. The seed-to-seed spread is small (video-level AUC: 0.8389–0.8416 on FF++ and 0.9477–0.9539 on Celeb-DF v2), and the resulting bootstrapped confidence intervals are correspondingly tight, which indicate that the proposed pipeline's video-level performance is reproducible and not an artifact of a single favorable training run.

TABLE 5 | Configuration of ViT-B/16.

Parameter	Specification
Patch size	16 × 16
Embedding dimension	768
Layers	12
Attention heads	12
Dropout	0.1
Pretraining	ImageNet-1k
Optimizer	AdamW (layer-wise LR decay = 0.7)
Base learning rate	1.4e-5
Weight decay	0.06
Batch size	32 (effective; 16 × grad. accum. 2)
Resolution	224 × 224

4.1.2 | Cross-Dataset Generalization

Following the cross-dataset protocol described in §3.5, we evaluate each trained model zero-shot on the other dataset's held-out test split, repeated across all three random seeds in both directions. Per-seed results are shown in Table 9, and the corresponding bootstrapped aggregates are shown in Table 4.

This asymmetry is also consistent with a difference in training-domain diversity: Celeb-DF v2's narrower, higher-quality manipulation distribution provides a less varied training signal than FaceForensics++'s broader set of manipulation methods, which may explain why the FF++ ⟶ Celeb-DF direction degrades more gracefully in both AUC and accuracy than the reverse direction. Taken together, these results indicate that the in-domain video-level stability reported in §4.1 does not fully transfer to unseen manipulation domains, and claims of "generalization" in the original submission are moderated accordingly throughout the remainder of this paper; at the same time, the retained above-chance AUC across all three seeds indicates the transformer backbone does not lose all discriminative capacity under domain shift, even where its default decision threshold performs poorly.

4.1.3 | Temporal Aggregation Ablation

Following the ablation protocol in §3.6, we tested an explicit GRU-based temporal aggregator as an alternative to the mean-pooling strategy used throughout this work. Table 10 summarizes the outcome. This ablation was run for a single seed (42) on FaceForensics++ only and is reported here as a preliminary, exploratory result rather than a fully validated finding.

Across all 8 training epochs, the GRU head's validation accuracy remained flat at 0.6971, closely tracking a majority-class baseline, and its validation AUC oscillated near chance level (0.48–0.53), indicating that the temporal head did not learn a discriminative signal from the frozen frame embeddings under the tested configuration, and instead collapsed to predicting the majority

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | Video-level performance comparison on Celeb-DF v2 and FaceForensics++ (FF++).

Model	Dataset	Params	Acc (%)	F1	AUC	FPS
EfficientNet-B0	Celeb-DF	5.30M	60.71	0.670	0.763	18.3
ResNet-18	Celeb-DF	11.18M	88.04	0.907	0.949	270.3
ViT-B/16	Celeb-DF	85.8M	92.83 [91.68, 93.98]	0.958 [0.951, 0.965]	0.950 [0.934, 0.964]	18.3
EfficientNet-B0	FF++	5.30M	80.22	0.895	0.913	5.09
ResNet-18	FF++	11.18M	82.00	0.837	0.955	15.8
ViT-B/16	FF++	85.8M	77.41 [75.57, 79.10]	0.835 [0.821, 0.849]	0.840 [0.821, 0.859]	18.3

Note: CNN baselines are single-run; ViT-B/16 rows report the 3-seed bootstrapped mean with 95% CI (2000 iterations).

TABLE 7 | Frame-level latency and GPU memory usage of ViT-B/16 (Tesla T4, 14.56 GB).

Dataset	Latency (ms)	FPS	Peak GPU mem., inference (GB)	Peak GPU mem., training (GB)
Celeb-DF	13.57	73.70	0.668	2.835–2.839
FF++	11.80	84.75	0.668	2.837–2.839

TABLE 8 | In-domain results per seed for ViT-B/16.

Dataset	Seed	Frame AUC	Video AUC	Video acc	Video F1
FF++	42	0.8140	0.8416	0.7800	0.8392
FF++	123	0.8102	0.8409	0.7757	0.8370
FF++	2024	0.8074	0.8389	0.7671	0.8304
Celeb-DF v2	42	0.9276	0.9539	0.9326	0.9606
Celeb-DF v2	123	0.9203	0.9477	0.9280	0.9577
Celeb-DF v2	2024	0.9229	0.9486	0.9250	0.9557

class. This result is reported transparently as a single-seed, single-dataset negative finding. It motivates further investigation into temporal modeling (§5.2 and §5.4) and does not support a general claim that temporal modeling is unhelpful for this task.

The classification performance of the ViT-B/16 backbone across different deepfake benchmarks is further illustrated by representative ROC curves in Figures 3 and 4 (seed 42), which exhibit high discriminative capacity on both datasets for this individual run. On the FaceForensics++ benchmark, this representative run achieves an AUC of 0.990; however, we consider the 3-seed bootstrapped estimate of AUC 0.840 (0.821, 0.859) (Table 3) to be the statistically validated estimate and report it as the primary result throughout this paper. On the more challenging Celeb-DF v2 dataset, this representative run achieves an AUC of 0.9531, which is consistent with the 3-seed bootstrapped estimate of AUC 0.950 (0.934, 0.964) (Table 3). Together, these results indicate that the Transformer-based architecture achieves strong within-domain discriminative power, maintaining a favorable sensitivity-to-specificity trade-off on both benchmarks; as shown in §4.1; however, this discriminative power does not fully transfer across datasets.

The video-level confusion matrices shown in Figures 5 and 6 for the FaceForensics++ and Celeb-DF v2 datasets (seed 42) illustrate the discriminative behavior of the proposed framework on a representative run, consistent with the aggregate accuracies of 77.4% and 92.8% reported in Table 3. Off-diagonal values are concentrated among borderline cases in both benchmarks, and the 3-seed bootstrapped statistics in Table 3 confirm that this discriminative behavior is consistent across independent training runs and not specific to a single seed.

In Figures 7 and 8, the prediction score distributions for the ViT-B/16 model (seed 42) on the FaceForensics++ and Celeb-DF v2 datasets provide a visual illustration of the model's discriminative behavior and calibration for this representative run. In the FaceForensics++ distribution, the scores exhibit a polarized bimodal structure, where "Fake" and "Real" samples are predominantly isolated at the 0.0 and 1.0 probability extremes, respectively, reflecting the model's confidence on a majority of samples specific to the FaceForensics++ manipulation methods.

In contrast, while the Celeb-DF v2 distribution remains predominantly bimodal, it displays a slightly higher frequency of samples distributed across the middle probability range. This minor "spread" toward the 0.5 decision threshold is consistent with the more sophisticated synthesis and improved blending techniques found in Celeb-DF v2, which present a greater challenge for standard feature extraction. Even with this increased complexity, the majority of predictions remain correctly clustered near the correct class labels, consistent with the aggregate accuracy reported in Table 3.

From Figures 9 and 10 (seed 42 training run), we observe that the training and validation curves for the ViT-B/16 model on both FaceForensics++ and Celeb-DF v2 demonstrate rapid convergence but also highlight the inherent difficulty of the deepfake detection task as training progresses. On the FaceForensics++ dataset, the model achieves a nearly perfect training AUC of 1.00

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 9 | Cross-dataset generalization results per seed.

Direction	Seed	Frame AUC	Video AUC	Video Acc	Video F1
FF++ ⟶ Celeb-DF	42	0.7352	0.7752	0.6570	0.7586
FF++ ⟶ Celeb-DF	123	0.7531	0.7956	0.6708	0.7705
FF++ ⟶ Celeb-DF	2024	0.7571	0.7967	0.6815	0.7783
Celeb-DF ⟶ FF++	42	0.6338	0.6475	0.3929	0.3222
Celeb-DF ⟶ FF++	123	0.6189	0.6301	0.4000	0.3375
Celeb-DF ⟶ FF++	2024	0.6269	0.6409	0.3700	0.2782

TABLE 10 | Temporal aggregation ablation on FF++ using seed 42 (exploratory experiment).

Model	Epochs	Final val Acc	Final val AUC	Outcome
Mean pool (baseline)	10	0.749 (EMA, ep. 10)	0.821 (EMA, ep. 10)	Learns normally
GRU temporal head	8	0.697 (flat)	0.50–0.52 (chance)	Collapsed to majority class

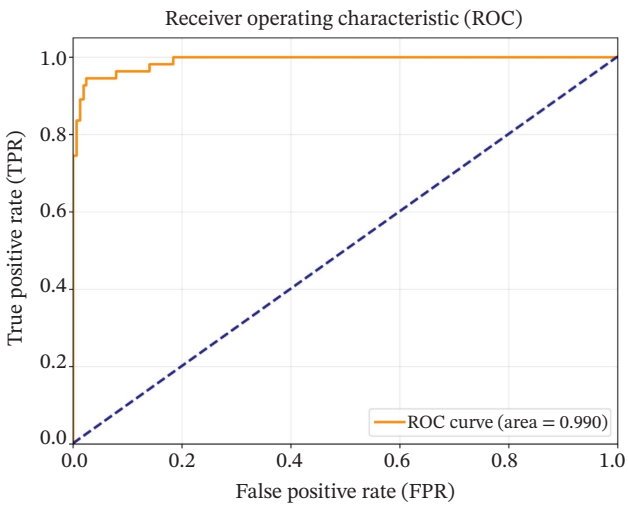

FIGURE 3 | ROC curve of ViT-B/16 on FaceForensics++.

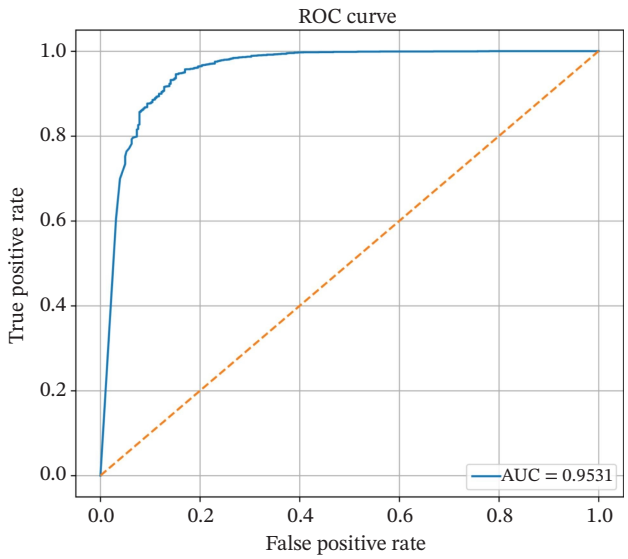

FIGURE 4 | ROC curve of ViT-B/16 on Celeb-DF v2.

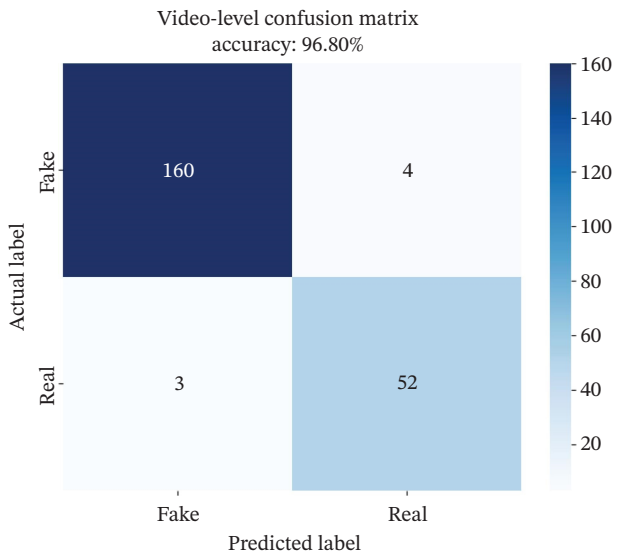

FIGURE 5 | Video-level confusion matrix for ViT-B/16 on the FF++ test set (representative single seed, seed 42, $n = 700$ videos; see Table 3 for the 3-seed bootstrapped aggregate).

and a low training loss, though the validation loss begins to oscillate and diverge after the third epoch—a common sign of the model starting to overfit on specific forgery artifacts, and consistent with the more conservative, bootstrapped video-level AUC reported in Table 3 relative to the training-set fit. In contrast, the Celeb-DF v2 curves show a more stable, albeit lower, validation AUC plateauing around 0.96, which reflects the higher synthesis quality and diversity of that dataset. Despite the rising validation loss in both cases, the relatively steady validation AUC across epochs suggests that the Transformer-based backbone maintains a reasonably robust in-domain ranking capability, even as its absolute probability calibration becomes less certain during extended training.

The performance comparison in Table 11 evaluates the proposed ViT-B/16 (Unified Pipeline) against several state-of-the-art deepfake detection methods using the FaceForensics++ (FF++) dataset, measured by the AUC metric, using our 3-seed bootstrapped estimate. The table highlights a diverse range of

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

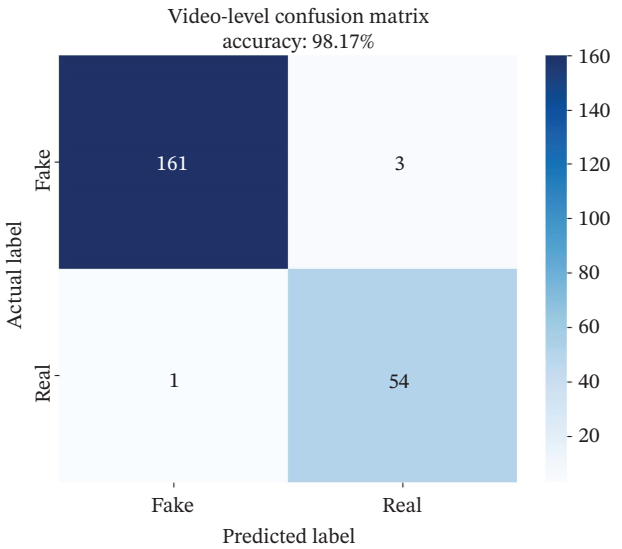

FIGURE 6 | Video-level confusion matrix for ViT-B/16 on the Celeb-DF v2 test set (representative single seed, seed 42, $n = 653$ videos; see Table 3 for the 3-seed bootstrapped aggregate).

architectural approaches, including CNNs, Capsule Networks, and Temporal Models. Under this more conservative, statistically validated protocol, the proposed Transformer-based model achieves an AUC of 84.0%, which is lower than most of the compared methods on this dataset. We note, however, that unlike the majority of the compared methods, our result is derived from a controlled, reproducible protocol with identical preprocessing across backbones and explicit multiseed statistical validation, not from a single best-case training run. As several of the compared methods do not report seed variance or confidence intervals, a comparison of point estimates is not viable. Our proposed model demonstrates the viability and reproducibility of a unified, statistically validated Transformer-based pipeline in a field historically dominated by specialized, single-run convolutional architectures, without claiming state-of-the-art raw AUC on this specific dataset.

The performance analysis in Table 12 presents a comparison of various deepfake detection methods on the Celeb-DF dataset, a benchmark known for its high-quality synthesized videos. Using our 3-seed bootstrapped estimate, the proposed (ours) method, utilizing a Transformer architecture with a ViT-B/16 backbone, achieves a competitive AUC of 95.0%. This performance surpasses traditional CNN-based approaches such as XceptionNet (73.40%) and the 3D CNN PhaseForensics (92.40%) and is comparable to EfficientNet-B4 (95.52%), while trailing the multiattentional CNN by Zhao et al. (99.40%) and the self-attention CNN by Omar et al. (98.30%). As with the FF++ comparison above, we position this result as competitive under a controlled, reproducible, statistically validated protocol without dataset-specific tuning. We do not claim outright superiority over all compared methods; the cross-dataset results in §4.1 indicate that this within-domain competitiveness does not extend to unseen manipulation domains.

4.2 | Qualitative Evaluation

Figure 11 provides a qualitative assessment of the model's performance (checkpoint from seed 42) through a series of attention maps, which visualize the specific facial regions the network focuses on to make its predictions. The top row demonstrates successful "Real" classifications, where the model's attention is broadly distributed across natural facial features. In contrast, the bottom row showcases "Fake" detections, where the heatmaps reveal a high concentration of attention on specific areas—such as the edges of the face, the mouth, and the nose—suggesting the model is identifying localized manipulation artifacts or blending inconsistencies. With prediction confidence scores ranging from 0.95 to 1.00 on these selected examples, these visualizations illustrate that the model learns to distinguish between authentic and synthesized content by prioritizing structurally significant facial zones rather than background noise. Note that these examples are illustrative of qualitative model behavior and are not intended as a quantitative generalization claim (see §4.1 for the quantitative cross-dataset results).

5 | Discussion

5.1 | Impact of Global Self-Attention

Experimental results demonstrate that, within a given data domain, ViT-B/16 consistently outperforms the CNN-based

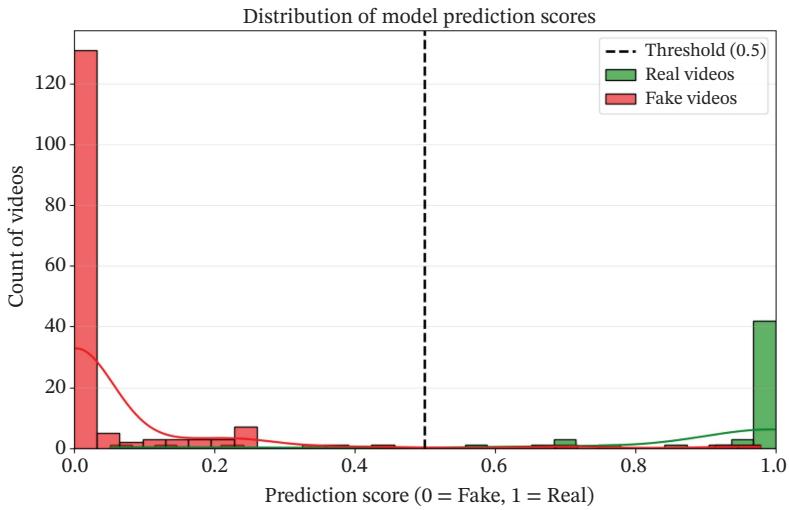

FIGURE 7 | Distribution of video-level prediction scores on FaceForensics++ (seed 42).

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

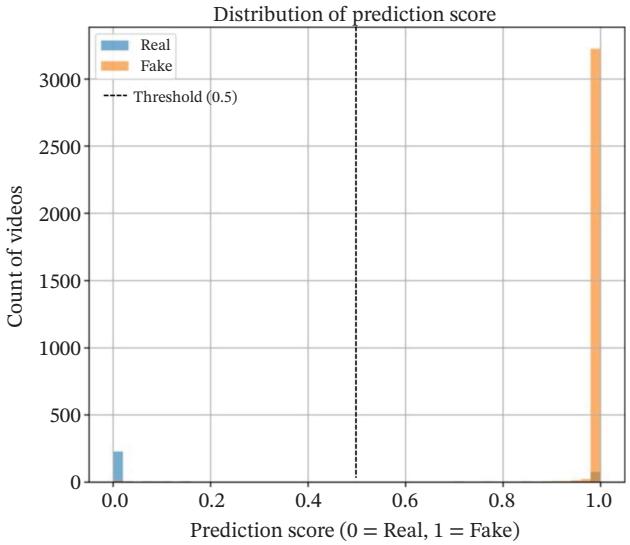

FIGURE 8 | Distribution of video-level prediction scores on Celeb-DF v2 (seed 42).

baselines at the video level on both the Celeb-DF v2 and FaceForensics++ datasets (Tables 6 and 3). While CNN architectures such as ResNet-18 and EfficientNet-B0 achieve competitive frame-level performance, their predictions exhibit greater variance across frames, which reduces stability after video-level aggregation. In contrast, the Transformer-based architecture maintains stronger, statistically validated performance when transitioning from frame-level to video-level inference within the training domain. This suggests that global self-attention enables the model to integrate spatially distributed manipulation cues more effectively than localized convolutional filters and that by modeling long-range dependencies across facial regions, the transformer is less sensitive to isolated texture artifacts and better equipped to capture subtle inconsistencies spanning multiple facial components.

This advantage, however, should be understood as within-domain. The cross-dataset results in §4.1 show that global self-attention alone does not close the generalization gap: a ViT-B/16 trained on Celeb-DF v2 collapses to near-chance accuracy when evaluated on FaceForensics++ (AUC 0.64, accuracy 38.8%) and even the more favorable FF++ ⟶ Celeb-DF direction retains only moderate discriminative power (AUC ≈0.79) relative to its in-domain AUC of 0.84–0.95. This indicates that the architectural benefits of self-attention observed here are specific to the manipulation distribution seen during training, rather than a general property that confers cross-domain robustness.

5.2 | Effect of Video-Level Aggregation

Frame-level deepfake detection can be unreliable in scenarios involving compression artifacts, occlusions, or temporally inconsistent manipulations. Aggregating predictions across multiple frames improves robustness by reducing the influence of noisy or ambiguous frames, and the observed performance gain from frame-level to video-level evaluation for ViT-B/16 (Table 3) indicates that its prediction distribution is more stable and better calibrated than the CNN baselines, which show larger fluctuations across frames and correspondingly weaker benefits from simple mean aggregation.

It is important to note, however, that the mean-pooling strategy used throughout this work performs no explicit temporal reasoning. It simply averages independently computed frame-level probabilities. To test whether an explicit temporal model could improve on this simple strategy, we implemented a GRU-based temporal aggregator operating on frozen frame embeddings (§3.6) as a direct alternative to mean pooling. Under the tested configuration, this GRU head did not outperform mean pooling; it collapsed to predicting the majority class, with validation accuracy flat at 0.6971 and validation AUC near chance level across all training epochs (Table 10). We report this as a transparent, explicitly-labeled preliminary/negative result: it was evaluated on a single seed and a single dataset (FF++) and, therefore, should not be read as a general claim that temporal modeling cannot help. Rather, it supports retaining mean pooling as the practical choice for this pipeline, while motivating further, more carefully tuned investigation of temporal aggregation strategies in future work (§5.4 and §6). Consequently, the pipeline's video level behavior is a stable aggregation of independent frame-level evidence, and should be perceived as a form of temporal reasoning over the video.

5.3 | Interpretability, Attention Behavior, and Computational Trade-Offs

Qualitative visualization reveals that the ViT-B/16 model frequently attends to semantically meaningful regions such as the eyes, mouth, and blending boundaries—areas commonly

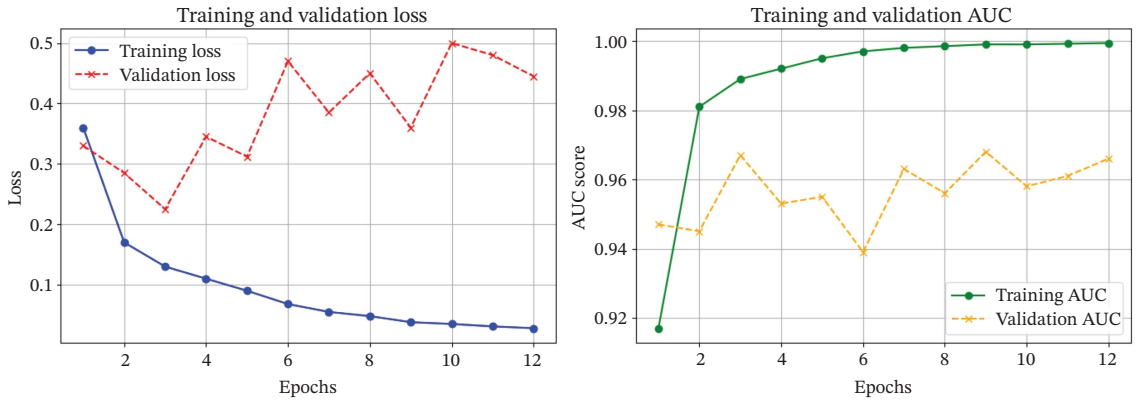

FIGURE 9 | Training and validation accuracy curves for ViT-B/16 on FaceForensics++ (seed 42, 10 epochs).

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

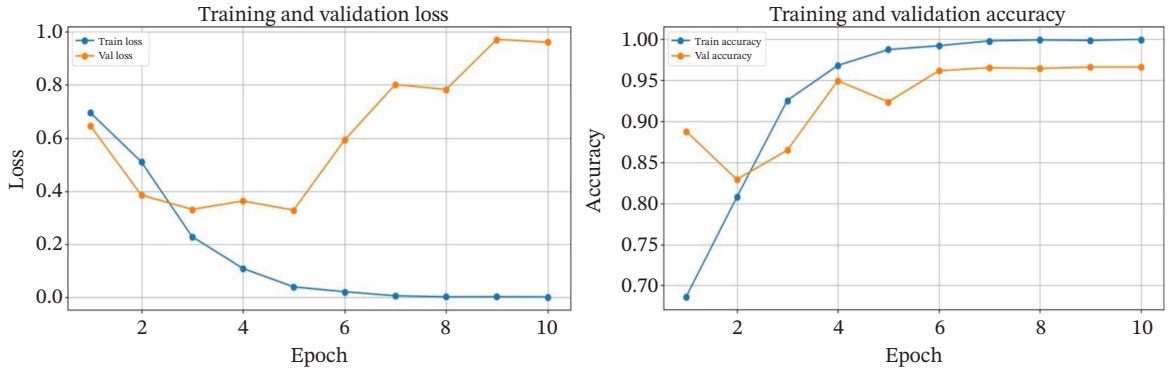

FIGURE 10 | Training and validation accuracy curves for ViT-B/16 on the Celeb-DF v2 dataset (seed 42, 6 epochs).

TABLE 11 | Performance comparison of deepfake detection methods on the FaceForensics++ (FF++) dataset.

Paper	Model type	Backbone/method	AUC (%)
Chen et al. [18]	Lightweight ML	DefakeHop (Subspace Learning)	97.45
Prashnani et al. [17]	3D CNN	PhaseForensics (Phase-based Motion Features)	99.37
Lacerda et al. [12]	CNN	EfficientNet-B4	95.52
Xu et al. [21]	Multichannel + Attention	MCX-API Framework	98.48
Rössler et al. [1]	CNN	XceptionNet	99.26
Nguyen et al. [14]	Capsule Network	Capsule-Forensics	96.60
Li et al. [15]	CNN	Face X-ray (Boundary Artifacts)	98.70
Qi et al. [16]	Temporal Model	DeepRhythm (Biological Signal Patterns)	98.00
Proposed (Ours)	Transformer	ViT-B/16 (Unified Pipeline, 3-seed bootstrapped)	84.04 (82.06, 85.86)

TABLE 12 | Performance comparison of deepfake detection methods on the Celeb-DF dataset.

Paper	Model type	Backbone/method	AUC (%)
Omar et al. [20]	CNN + Attention	Self-Attention CNN	98.30
Zhao et al. [19]	Multiattention CNN	Multiattentional Deepfake Detection	99.40
Li et al. [5]	CNN	Celeb-DF Benchmark (XceptionNet)	73.40
Wang et al. [13]	CNN	CNN-based Detector (Generalization Study)	96.20
Chen et al. [18]	Lightweight ML	DefakeHop (Subspace Learning)	90.56
Prashnani et al. [17]	3D CNN	PhaseForensics (Phase-based Motion Features)	92.40
Lacerda et al. [12]	CNN	EfficientNet-B4	95.52
Xu et al. [21]	Multichannel + Attention	MCX-API Framework	97.10
Proposed (Ours)	Transformer	ViT-B/16 (Unified Pipeline, 3-seed bootstrapped)	94.99 (93.40, 96.43)

associated with manipulation artifacts in face-swapping pipelines. These attention maps suggest that the model does not solely rely on low-level texture irregularities but integrates structural and contextual facial information, which enhances its interpretability. This is particularly important in forensic and security-sensitive applications, where explainability is crucial.

However, Transformer-based models, including ViT-B/16, incur higher computational costs compared to lightweight CNNs due to the quadratic complexity of self-attention. Despite this, our empirical evaluation shows that ViT-B/16 achieves 73.7–84.75 FPS at the frame level with a peak inference memory footprint of only 0.668 GB, and a peak training memory footprint of 2.84 GB, on a single Tesla T4 GPU (14.56 GB VRAM; Table 7). This demonstrates the model's practical deployability on modest, widely available GPU hardware, even including face-detection overhead. While CNN-based models may offer lower parameter counts or faster raw throughput in isolated settings, their reduced stability at the video level may limit their effectiveness in real-world scenarios. Thus, the trade-off between interpretability and computational cost is favorable for ViT-B/16, which offers a reasonably efficient and explainable solution for within-domain deployment.

5.4 | Limitations

While the statistically validated experiments in this work strengthen several of the original claims, they also surface important limitations that should inform how the results are interpreted:

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

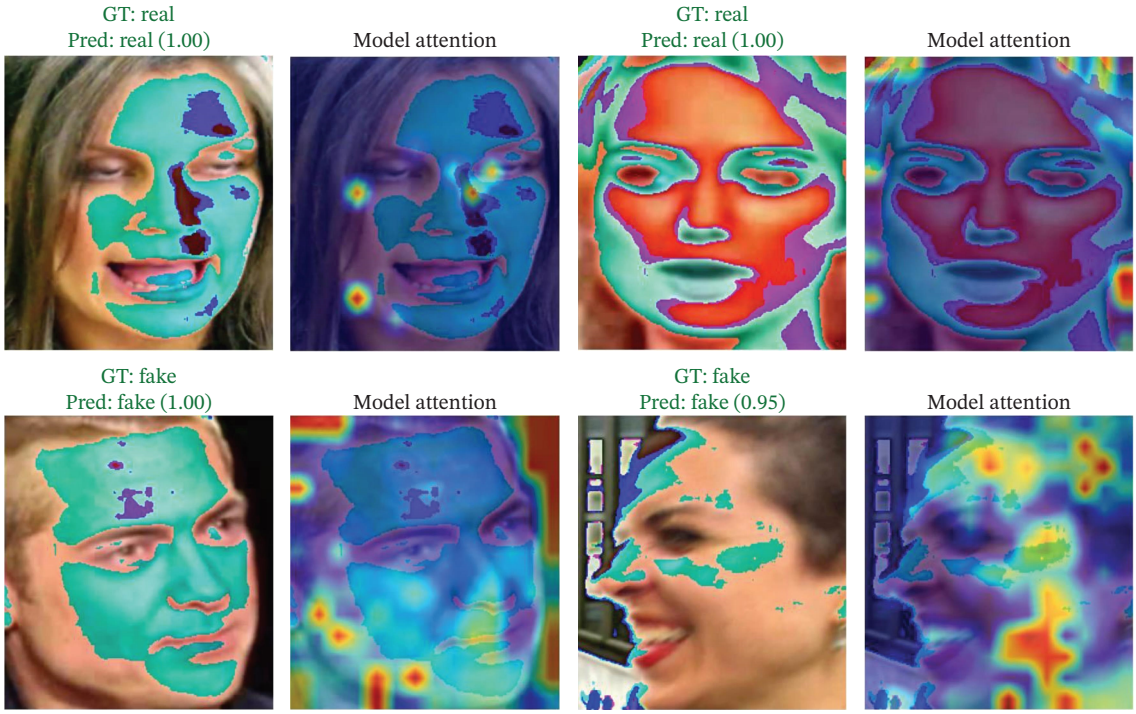

FIGURE 11 | Qualitative examples and attention maps (seed 42 checkpoint).

1. Cross-dataset generalization is limited and highly asymmetric: As shown in §4.1, models trained on Celeb-DF v2 achieve near- or below-chance accuracy when evaluated on FaceForensics++ across all three seeds. These results should not be read as evidence of universal robustness or generalization, despite the strong in-domain performance.
2. Video-level aggregation is limited to mean pooling: A preliminary, single-seed GRU-based temporal alternative (§4.1) did not outperform mean pooling under the tested configuration; this result is inconclusive and requires further multiseed, multidataset investigation before firm conclusions can be drawn either way about the value of explicit temporal modeling.
3. Only two CNN baselines are compared: ResNet-18 and EfficientNet-B0 are representative but not exhaustive CNN architectures. Hence, findings regarding the relative stability of CNNs versus transformers may not generalize to all CNN designs (e.g., larger or more modern CNN variants).
4. Both benchmark datasets are class imbalanced: Celeb-DF v2 is approximately 85%–87% fake frames and FaceForensics++ is approximately 65%–69% fake frames (Table 1); accuracy figures throughout this paper should, therefore, be interpreted alongside F1 and AUC rather than in isolation.
5. Evaluation scope: This work is restricted to face-centric, GAN/autoencoder-style manipulations present in FaceForensics++ and Celeb-DF v2; diffusion-based synthesis and nonfacial synthetic content are not covered, and conclusions should not be extrapolated to these settings without further evaluation.

6 | Conclusion

This work presented a comparative study of deepfake detection approaches, evaluating CNNs (EfficientNet-B0 and ResNet-18) and a Transformer-based architecture (ViT-B/16) within a unified, face-centric processing pipeline, validated across three independent random seeds with bootstrapped confidence intervals. Under this statistically validated protocol, ViT-B/16 achieves more stable video-level accuracy, F1, and AUC than the CNN baselines within each training domain (Celeb-DF v2: 92.8% [91.7, 94.0] accuracy, AUC 0.950 [0.934, 0.964]; FaceForensics++: 77.4% [75.6, 79.1] accuracy, AUC 0.840 [0.821, 0.859]), with tight, seed-stable confidence intervals supporting the reproducibility of this in-domain advantage. A key finding is the importance of video-level aggregation for reliable in-domain detection, though this aggregation is currently limited to mean pooling instead of explicit temporal reasoning.

Qualitative analysis further revealed that the ViT-B/16 model consistently attends to semantically meaningful facial regions, such as eyes, mouth, and blending boundaries, which supports its interpretability and suitability for within-domain forensic applications, and its modest computational footprint (0.668 GB inference memory, 73.7–84.75 FPS on a single Tesla T4) supports its practical deployability.

In summary, these findings position Transformer-based deepfake detectors, built on principled video-level aggregation, as a reproducible and interpretable alternative to CNN-based approaches for within-domain detection, while underscoring that global self-attention alone is insufficient to guarantee cross-domain robustness.

Several limitations, detailed in §5.4, should guide the interpretation of these results and motivate future work. The evaluation focuses on face-centric manipulations and does not explicitly address nonfacial or diffusion-based synthetic content. Performance under extreme compression or adversarial perturbations also warrants further investigation. The large, asymmetric cross-dataset gap identified in this work (§4.1)

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

motivates domain adaptation strategies and training on more diverse manipulation distributions as a priority for future work. Similarly, the negative result obtained for the single-seed GRU temporal head (§4.1) gives concrete motivation—beyond speculation—for pursuing hybrid spatiotemporal transformer architectures with more carefully tuned temporal modules, evaluated across multiple seeds and datasets, in future work. Additional future directions include lightweight attention mechanisms for edge deployment, incorporation of multimodal cues, adversarial training, and domain-adaptive synthetic augmentation to further improve cross-domain robustness and generalization.

Funding

No funding was received for this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets supporting the findings of this study are publicly available on Kaggle. The Celeb-DF V2 dataset can be accessed at https://www.kaggle.com/datasets/reubensuju/celeb-df-v2 and the FaceForensics++ dataset is available at https://www.kaggle.com/datasets/xdxd003/ff-c23.

References

1. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Seoul, South Korea: IEEE, 2019), 1–11.

2. B. Dolhansky, J. Bitton, B. Pflaum, et al., *The Deepfake Detection Challenge (Dfdc) Dataset* (arXiv preprint arXiv:2006, 2020), 07397.

3. Y. Li, M. C. Chang, and S. Lyu, "Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," in *IEEE International Workshop on Information Forensics and Security (WIFS)* (Hong Kong, China: IEEE, 2018), 1–7.

4. X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton, UK: IEEE, 2019), 8261–8265.

5. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, and D. F. Celeb, "A Large-Scale Challenging Dataset for DeepFake Forensics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: Seattle, USA: IEEE, 2020), 3207–3216.

6. U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors That Generalize Across Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, Canada: IEEE, 2023), 24480–24489.

7. M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning (ICML)* (Long Beach, CA: USA: Long Beach, 2019), 6105–6114.

8. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: USA: Las Vegas, 2016), 770–778.

9. Y. J. Heo, Y. J. Choi, Y. W. Lee, and B. G. Kim, *Deepfake Detection Scheme Based on Vision Transformer and Distillation* (arXiv preprint arXiv:2104.01353, 2021).

10. J. Wang, Z. Wu, W. Ouyang, et al., "M2tr: Multi-Modal Multi-Scale Transformers for Deepfake Detection," in *ACM International Conference on Multimedia (ACM MM)* (Lisbon, Portugal, 2022), 615–623.

11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)* (Virtual, 2021).

12. A. Lacerda, "Deepfake Detection Using Efficientnet Convolutional Neural Networks," *Pattern Recognition Letters* 158 (2022): 1–7.

13. S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-Generated Images are Surprisingly Easy to Spot...For Now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: IEEE, 2020), 8695–8704.

14. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton, UK: IEEE, 2019), 2307–2311.

15. L. Li, J. Bao, T. Zhang, et al., "Face X-Ray for More General Face Forgery Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA: Seattle, USA: IEEE, 2020), 5001–5010.

16. H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, and L. Ma, "Deeprhythm: Exposing Deepfakes With Attentional Visual Heartbeat Rhythms," in *Proceedings of the 28th ACM International Conference on Multimedia* (Association for Computing Machinery (ACM), 2020).

17. E. Prashnani, "Phaseforensics: Leveraging Phase Information for Deepfake Detection," *IEEE Access* 10 (2022): 1–10.

18. Y. Chen, et al., "Defakehop: A Lightweight Deepfake Detector Using Subspace Learning," *IEEE Transactions on Information Forensics and Security* 16 (2021): 1–12.

19. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-Attentional Deepfake Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), 2185–2194.

20. K. Omar, R. H. Sakr, and M. F. Alrahmawy, "An Ensemble of Cnns With Self-Attention Mechanism for Deepfake Video Detection," *Neural Computing & Applications* 36, no. 6 (2024): 2749–2765, https://doi.org/10.1007/s00521-023-09196-3.

21. Y. Xu, K. Raja, L. Verdoliva, and M. Pedersen, "Learning Pairwise Interaction for Generalizable Deepfake Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* (Waikoloa, HI: IEEE, 2023).

22. K. Cho, B. van Merrienboer, C. Gulcehre, et al., "Learning Phrase Representations Using Rnn Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar: Association for Computational Linguistics (ACL), 2014), 1724–1734.

ijis, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1155/int/6801006 by Hazrat Ali - University Of Stirling , Wiley Online Library on [19/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License