WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Transformer-Based Multi-Modal Architecture for Real-Time Deepfake Detection in Streaming Video

Mr. Manish Jain

Assistant Professor
Jagannath University, Jaipur

ABSTRACT :

The rapid diffusion of generative deep learning models—specifically Diffusion-based video synthesis, Neural Radiance Fields (NeRFs), and advanced voice cloning autoencoders—has made synthetic media creation accessible to non-experts. While these techniques drive innovations in entertainment and virtual reality, they also enable severe threats, including identity fraud, automated disinformation campaigns, and financial manipulation via real-time streaming video feeds. Existing deepfake detection methods largely rely on single-modality unimodal classifiers (e.g., frame-by-frame Spatial Convolutional Neural Networks) that fail under dynamic streaming conditions due to frame drops, adaptive bitrate compression artifacts, and low latency budgets. This paper introduces an end-to-end Transformer-based Multi-Modal Architecture designed for real-time deepfake detection in live streaming video pipelines. By fusing spatial-temporal visual features (derived from a lightweight Video Vision Transformer, ViViT) with acoustic-phonetic speech representations (extracted via a self-supervised Wav2Vec 2.0 backbone) through a Cross-Modal Attention Bridge (CMAB), our model captures subtle audio-visual desynchronizations and lip-phoneme misalignments. To achieve real-time streaming throughput (>60 FPS at <45 ms end-to-end latency), we integrate TensorRT graph optimizations, structured token pruning, and a sliding-window temporal memory buffer. Benchmark evaluations across benchmark datasets (DFDC, FaceForensics++, and AV-Deepfake1M) demonstrate a superior Area Under the Curve (AUC) of 98.4% and an Equal Error Rate (EER) of 2.1%, outperforming unimodal and late-fusion baselines while operating efficiently on edge GPU infrastructure.

1. Introduction & Problem Formulation

Synthetic media generation, colloquially termed 'deepfakes', has transitioned from an academic curiosity to a high-impact cybersecurity vulnerability. Driven by advancements in Generative Adversarial Networks (GANs), Denoising Diffusion Probabilistic Models (DDPMs), and neural audio synthesis, modern deepfake generation pipelines can produce photorealistic video and natural-sounding voice clones in real time. When deployed over live streaming platforms, social media webcasts, and video conferencing channels, manipulated media threatens public trust, financial markets, and enterprise identity security.

Real-time streaming video poses three distinct computational and algorithmic challenges for automated deepfake detection:

1. Adaptive Bitrate and Compression Noise: Streaming protocols (e.g., WebRTC, RTMP) dynamically adjust video resolution, frame rates, and compression levels based on network throughput. Spatial compression artifacts (H.264/H.265 intra-frame macroblocking) often destroy subtle high-frequency spatial cues relied upon by conventional CNN detectors.

2. Latency and Throughput Constraints: Live forensic analysis requires strict frame processing deadlines (<50 ms total pipeline latency) to enable real-time flagging or automated stream termination, rendering heavy 3D-CNNs or multi-stage ensembles computationally prohibitive.

3. Cross-Modal Desynchronization: Advanced deepfake generators synthesize highly convincing spatial facial textures but frequently fail to maintain cross-modal coherence—such as precise synchrony between acoustic phonemes (e.g., bilabial plosives /p/, /b/, /m/) and visual visemes (lip geometry).

To resolve these challenges, this paper presents a novel multi-modal Transformer architecture explicitly engineered for streaming video streams. By jointly processing synchronized visual streams and acoustic waveforms through cross-attention mechanisms, the proposed framework detects subtle cross-modal incoherencies while maintaining real-time edge processing speeds.

2. Related Work & Detection Paradigms

Deepfake detection methodologies are broadly categorized into three technical paradigms:

2.1 Unimodal Visual Detectors

Early detection systems relied on frame-level Convolutional Neural Networks (EfficientNet, ResNet) trained to identify spatial artifacts, such as warping anomalies, eye-blinking irregularities, or boundary inconsistencies around facial masks. Later architectures introduced 3D-CNNs (e.g., Inflated 3D ConvNets) and Vision Transformers (ViT) to capture temporal continuity across contiguous frames. However, visual-only detectors remain vulnerable to adversarial perturbations and severe lossy compression common in live streaming pipelines.

2.2 Audio Deepfake and Voice Clone Detection

Acoustic forensics focuses on identifying voice synthesis artifacts by analyzing spectral features, mel-frequency cepstral coefficients (MFCCs), or raw audio waveforms using self-supervised models such as Wav2Vec 2.0 and HuBERT. While effective against isolated voice cloning, audio-only detectors cannot determine whether a genuine audio track has been swapped onto an incompatible visual face feed.

2.3 Multi-Modal Audio-Visual Fusion

Recent studies demonstrate that combining audio and visual modalities significantly improves detection robustness. Existing multi-modal approaches primarily rely on late fusion (averaging predictions from independent visual and audio classifiers) or early concatenation. Late fusion fails to capture fine-grained temporal alignment between phonemes and visemes, whereas early concatenation creates high-dimensional vector spaces that are difficult to optimize. Our work addresses these gaps by introducing an intermediate cross-attention transformer fusion layer specifically optimized for real-time stream processing.

3. System Architecture & Multi-Modal Fusion

The proposed system architecture processes synchronized audio-visual chunks from a live video stream through four key stages: (1) Stream Ingestion & Preprocessing, (2) Spatial-Temporal Visual Encoder, (3) Acoustic-Phonetic Audio Encoder, and (4) Cross-Modal Attention Bridge (CMAB).

3.1 Visual Stream Processing (ViViT Backbone)

Given a video stream chunk V in R^{T x H x W x C} spanning T = 16 frames, facial region-of-interest (ROI) bounding boxes are tracked using a lightweight face detector. The ROI sequences are split into non-overlapping spatial-temporal 3D tubelets of dimension p_t x p_h x p_w. These tubelets are linearly projected into a sequence of visual tokens Z_v in R^{N_v x d_v}, augmented with spatial and temporal positional embeddings, and passed through a lightweight Video Vision Transformer (ViViT) encoder.

3.2 Audio Stream Processing (Wav2Vec 2.0 Backbone)

The synchronized audio segment A corresponding to the 16-frame window is sampled at 16 kHz. Raw audio waveforms pass through a multi-layer 1D convolutional network to extract local acoustic latent vectors, which are subsequently processed by a Transformer encoder pre-trained via self-supervised contrastive learning (Wav2Vec 2.0). This yields acoustic sequence tokens Z_a in R^{N_a x d_a} encoding rich phonetic structures.

3.3 Cross-Modal Attention Bridge (CMAB)

To capture fine-grained inter-modality relationships, the visual tokens Z_v and audio tokens Z_a are projected into a shared latent space of dimension d_m. The projected tokens enter the Cross-Modal Attention Bridge (CMAB), where visual tokens act as Queries (Q_v) while audio tokens serve as Keys (K_a) and Values (V_a), and vice versa:

Equation (1): Attention(Q_v, K_a, V_a) = Softmax\left(\frac{Q_v K_a^T}{\sqrt{d_k}}\right) V_a

This bi-directional cross-attention mechanism explicitly measures how well visual lip movements match the underlying speech phonemes, producing a unified multi-modal feature vector Z_f that highlights audio-visual misalignment anomalies.

4. Real-Time Optimization for Streaming Inference

Deploying transformer models on live video streams requires aggressive optimization to satisfy strict real-time operational boundaries without sacrificing detection accuracy.

4.1 Sliding-Window Temporal Memory Buffer

Instead of re-computing full transformer attention maps for every incoming frame chunk, we implement a Sliding-Window Temporal Memory Buffer. The buffer retains key and value representations (KV-cache) from previous frames, allowing the model to update predictions incrementally for new incoming frames. This reduces computational overhead by 65% compared to non-overlapping batch processing.

4.2 Token Pruning and TensorRT Graph Compilation

To eliminate redundant spatial background tokens, an adaptive token selection module computes attention score variances and prunes non-informative background tokens prior to the cross-attention fusion stage. Furthermore, the complete multi-modal graph is compiled using NVIDIA TensorRT, enabling FP16 mixed-precision acceleration, kernel auto-tuning, and dynamic memory allocation on edge GPUs.

5. Experimental Setup & Comparative Benchmarking

We evaluate our proposed architecture across three benchmark deepfake datasets: FaceForensics++ (FF++), Deepfake Detection Challenge (DFDC), and AV-Deepfake1M. Performance is measured using Area Under the ROC Curve (AUC), Equal Error Rate (EER), and Frame Processing Speed (FPS).

Architecture / Baseline	Modality	FF++ (AUC %)	DFDC (AUC %)	EER (%)	Inference Speed (FPS)
EfficientNet-B4 + Frame CNN	Visual Only	92.4%	81.2%	8.5%	45 FPS
3D-ResNet50 Temporal	Visual Only	95.1%	84.6%	6.2%	28 FPS
Late Fusion (ViT + Wav2Vec)	Audio-Visual	96.8%	89.3%	4.1%	38 FPS
Proposed CMAB Transformer (FP16)	Audio-Visual Cross-Attn	98.4%	93.7%	2.1%	68 FPS

6. Ablation Studies & Robustness Analysis

To validate individual components, systematic ablation experiments were conducted:

1. Impact of Cross-Modal Bridge: Removing the Cross-Modal Attention Bridge and replacing it with simple vector concatenation reduced the DFDC AUC score by 4.4%, demonstrating the importance of explicit cross-modal attention for capturing lip-phoneme misalignments.

2. Robustness to Bitrate Compression: Subjecting streaming inputs to simulated packet loss (5% to 15%) and H.264 compression bitrate throttling showed that while visual-only models suffered a 12.8% drop in AUC, our multi-modal model maintained high accuracy (drop < 2.1%) by relying on uncorrupted audio-phoneme streams.

3. Throughput and Latency Optimization: Enabling token pruning and TensorRT quantization reduced end-to-end processing latency from 82 ms per chunk to 32 ms per chunk, enabling throughput (>60 FPS) on an NVIDIA RTX 4090 / Jetson Orin setup.

7. Conclusion & Future Directions

This paper introduced a Transformer-based multi-modal architecture tailored for real-time deepfake detection in streaming video applications. By integrating spatial-temporal visual tokens with acoustic speech representations via a Cross-Modal Attention Bridge (CMAB), the framework effectively identifies audio-visual desynchronizations that unimodal models fail to detect. Combined with token pruning and TensorRT optimizations, the model achieves state-of-the-art accuracy (98.4% AUC on FF++) while maintaining low-latency inference (>60 FPS).

Future research will expand the architecture to analyze spatial multi-speaker audio-visual feeds, evaluate zero-shot detection capabilities against novel diffusion-based neural avatars, and implement lightweight model compression for deployment on consumer mobile devices.

References

[1] R. Tolosana et al., "Deepfakes and beyond: A survey of face manipulation and detection," Information Fusion, vol. 64, pp. 131–148, 2020.
[2] Y. Arnab et al., "Vivit: A video vision transformer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 6836–6846.
[3] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," Adv. Neural Inf. Process. Syst. (NeURIPS), vol. 33, pp. 12449–12460, 2020.
[4] T. Mittal et al., "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in Proc. ACM Int. Conf. Multimedia, 2020, pp. 2823–2832.
[5] H. Zhou et al., "Pose-controllable talking face generation by implicitly modularized audio-visual representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 8681–8696, 2021.