International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

# A Multimodal Deep Learning Framework for Robust Deepfake Detection Using Audio-Visual and Temporal Features

**Vaibhav Singh[1], Dr. Vidhya Kshirsagar[2]**

[1]*Student, Department of Master of Computer Applications (MCA), D.Y. Patil Deemed to be University, School of Humanities and Sciences, Navi Mumbai, Maharashtra, India*
[2]*Assistant Professor, Department of Master of Computer Applications (MCA), D.Y. Patil Deemed to be University, School of Humanities and Sciences, Navi Mumbai, Maharashtra, India*

---

**Abstract -** *The rapid proliferation of generative artificial intelligence, including latent diffusion architectures, neural audio synthesizers, and Generative Adversarial Networks (GANs), has democratized the synthesis of hyper-realistic face swaps, talking heads, and cloned voices. While existing automated forensic detectors achieve high accuracy on uncompressed laboratory benchmarks, they rely predominantly on unimodal visual representations from isolated frames and exhibit severe vulnerability under lossy video compression, asymmetric audio-visual manipulations, and cross-generator distribution shifts. To overcome these limitations, this study presents an end-to-end multimodal deep learning framework that couples fine-grained spatial facial representations, acoustic spectral dynamics, and bidirectional cross-modal temporal dependencies. Facial regions are localized and aligned via RetinaFace with five-point affine transformation and encoded using a Vision Transformer (ViT-B/16) backbone. Simultaneously, synchronized 16 kHz audio tracks are converted into 128-channel log-Mel spectrograms using a 25 ms Hanning window and encoded via a ResNet-18 network. A bidirectional scaled dot-product cross-modal attention module queries acoustic phoneme representations using visual viseme embeddings, and the fused sequences are modeled across 32 frames through a four-layer temporal Transformer encoder trained under a joint binary cross-entropy and contrastive synchronization objective. Evaluated across FakeAVCeleb, FaceForensics++ (Raw, c23, and c40), and the Deepfake Detection Challenge (DFDC), the proposed framework achieves 96.8% ROC-AUC, 93.8% F1-score, and 93.8% accuracy on FakeAVCeleb, outperforming the ViT-B/16 visual baseline by 5.6% ROC-AUC and the ResNet-18 acoustic baseline by 11.4% ROC-AUC. Under heavy H.264 c40 compression, it retains 88.9% ROC-AUC (only 7.1% accuracy reduction versus 24.3% unimodal collapse), and achieves 84.7% zero-shot ROC-AUC on unseen DFDC manipulations.*

**Key Words: Deepfake Detection, Multimodal Fusion, Audio-Visual Synchronization, Vision Transformers (ViT-B/16), Temporal Consistency, Phoneme-Viseme Alignment, Digital Media Forensics, H.264 Compression Resilience.**

## 1. INTRODUCTION

Recent advancements in generative deep learning have fundamentally transformed the landscape of synthetic media production, drastically lowering the technical and computational barriers required to fabricate hyper-realistic audiovisual recordings. Contemporary manipulation pipelines span multiple algorithmic families, ranging from identity-swapping frameworks such as DeepFaceLab and SimSwap to audio-driven facial reenactment models like Wav2Lip [13] and SadTalker, alongside emerging latent diffusion video generators and zero-shot neural voice cloning vocoders. As documented in foundational surveys of synthetic media forensics [7], these architectures generate high-definition video sequences in which facial textures, illumination gradients, and vocal prosody are virtually indistinguishable from authentic recordings to unaided human observers. While generative audiovisual synthesis offers substantial utility in motion picture post-production, virtual telepresence, multilingual dubbing, and assistive communication technologies, its malicious weaponization poses severe threats to democratic elections, judicial evidence integrity, corporate financial authorization protocols, and individual privacy.

Historically, the foundational paradigm of automated deepfake forensics has focused primarily on identifying localized spatial anomalies within individual video frames. First-generation forensic classifiers targeted low-level pixel and geometric irregularities, including inconsistent corneal and pupillary specular reflections, abnormal physiological eye-blinking intervals, affine warping seams along the facial periphery, and chromatic aberration discrepancies between spliced foreground masks and background scenes [8]. However, as generative architectures evolved, synthesis pipelines began incorporating adversarial perceptual loss functions, multi-scale Poisson blending, and frequency-domain boundary smoothers directly into their training objectives. These generative refinements effectively suppress static spatial footprints, enabling modern deepfakes to exhibit pristine frame-level visual fidelity while concealing subtle structural anomalies across time and across sensory modalities.

A second critical vulnerability of purely spatial visual forensics lies in how digital media is distributed in real-world operational environments. Before reaching end users or forensic analysts, videos uploaded to commercial social media platforms, cloud messaging applications, and streaming services are systematically subjected to aggressive lossy transcoding using H.264/MPEG-4 AVC or H.265/HEVC codecs with high Constant Rate Factors (CRF). During block-based discrete cosine transform (DCT) quantization and chroma subsampling, high-frequency spatial details are selectively discarded to minimize bandwidth consumption. Because unimodal spatial detectors rely heavily on high-frequency pixel-level blending traces, heavy channel compression destroys their primary forensic signal, causing severe accuracy degradation precisely where automated verification is needed most.

A third operational challenge stems from the multimodal nature of modern media manipulation. Bad actors rarely restrict themselves to modifying a single sensory channel. A deceptive video may combine a pristine, unaltered facial recording with a cloned, synthetic audio track generated via neural text-to-speech (RealVideo-FakeAudio), or conversely, pair an authentic human voice recording with a lip-reenacted synthetic video stream (FakeVideo-RealAudio), or synthesize both modalities simultaneously (FakeVideo-FakeAudio) [9]. Unimodal visual detectors are inherently blind to audio-only voice cloning when the visual stream remains untouched or minimally reenacted, whereas unimodal acoustic classifiers fail completely when genuine speech is used to drive a swapped or reenacted face. Consequently, any forensic architecture that evaluates vision and audio in isolation cannot provide comprehensive protection against asymmetric multimodal deception.

To construct durable, compression-resilient forensic defenses, automated verification systems must transition from isolated static frame analysis to unified spatiotemporal and audio-visual modeling. Natural human speech production is governed by strict biomechanical and anatomical constraints: the acoustic articulation of specific phonetic units (phonemes such as bilabial plosives /p/, /b/, /m/, labiodental fricatives /f/, /v/, and rounded vowels) is physically produced by coordinated muscular deformations of the lips, jaw, tongue, and lower vocal tract (visemes) [6]. When a synthetic face is spliced onto a target video, or when synthetic speech is coupled with an independently generated facial animation sequence, micro-temporal phase shifts, incomplete labial closures, and phoneme-viseme desynchronization inevitably emerge across consecutive frames. Crucially, unlike high-frequency pixel blending noise, these macroscopic articulatory trajectories and low-to-mid-frequency acoustic spectral envelopes survive aggressive H.264/H.265 video compression.

This paper presents a comprehensive study and implementation of A Multimodal Deep Learning Framework for Robust Deepfake Detection Using Audio-Visual and Temporal Features. Rather than treating video frames and audio waveforms as independent classification inputs, the proposed framework models them as tightly coupled, co-evolving physical signals within an end-to-end deep neural architecture. In the visual stream, every input video clip is processed through a single-stage RetinaFace localization network [12] with five-point facial landmark affine alignment, normalized to 224×224×3 RGB tensors, and partitioned into 16×16 spatial patches for encoding via a Vision Transformer (ViT-B/16) backbone [11]. In parallel, the acoustic stream extracts the raw audio waveform at a 16 kHz sampling rate, computes a 128-bin log-Mel spectrogram using a 25 ms Hanning window with a 10 ms hop length, and extracts temporally aligned acoustic embeddings via a ResNet-18 convolutional encoder.

To capture fine-grained cross-modal correspondence, the framework introduces a Bidirectional Cross-Modal Attention module in which spatial visual embeddings dynamically query acoustic keys and values via scaled dot-product attention at the frame level. The resulting cross-attended representation sequence is concatenated with residual visual tokens and passed into an L-layer Temporal Transformer Encoder to model long-range inter-frame articulatory dynamics across T = 32 consecutive frames. The entire architecture is jointly optimized using standard Binary Cross-Entropy ($L_{BCE}$) paired with a contrastive audio-visual synchronization loss ($L_{sync}$) that explicitly penalizes cross-modal feature divergence on authentic recordings. Through extensive benchmarking on FakeAVCeleb, FaceForensics++ (across Raw, c23, and c40 compression tiers), and the Deepfake Detection Challenge (DFDC) dataset, this study demonstrates how deep multimodal temporal fusion establishes superior accuracy, compression robustness, and zero-shot cross-generator generalization.

## 2. PROBLEM STATEMENT

Current automated deepfake detection methodologies remain heavily skewed toward unimodal spatial visual analysis, rendering deployed forensic pipelines vulnerable to three systemic operational failures when confronted with real-world synthetic media:

### 2.1 Compression Fragility under Lossy Transmission Channels

Commercial social media ecosystems, video-sharing networks, and encrypted messaging platforms routinely transcode user-uploaded videos using aggressive lossy compression standards principally H.264/MPEG-4 AVC and H.265/HEVC configured with high Constant Rate

Factors (such as CRF 23 for standard streaming and CRF 40 for low-bandwidth transmission). During macroblock partitioning and frequency-domain quantization, high-frequency spatial coefficients are zeroed out or coarsely approximated. Because conventional CNN and Vision Transformer detectors depend on microscopic pixel-level blending boundaries, checkerboard upsampling artifacts, and local color-space inconsistencies, heavy compression erases their primary decision cues while simultaneously introducing block-boundary compression artifacts into authentic videos, leading to severe false-positive and false-negative spikes.

### 2.2 Generalization Failure across Unseen Generative Paradigms

Unimodal visual detectors trained on conventional GAN-based identity-swapping datasets (such as Deepfakes or FaceSwap in FaceForensics++) experience catastrophic performance degradation when evaluated zero-shot against unseen manipulation paradigms, including audio-driven talking-head synthesizers (Wav2Lip, SadTalker) and latent diffusion reenactment pipelines. Spatial feature extractors rapidly overfit to generator-specific upsampling fingerprints and training-set color distributions rather than learning universal, generator-agnostic invariants of human facial articulation.

### 2.3 Acoustic-Visual Asymmetry and Cross-Modal Blindness

Modern multimodal forgeries selectively manipulate either the visual track, the acoustic track, or both. Unimodal acoustic forensics models fail completely when an authentic, unedited voice recording is paired with a synthetic identity-swapped video stream (FakeVideo-RealAudio). Conversely, unimodal visual classifiers fail to flag recordings in which an authentic video of a speaker is paired with a cloned audio track accompanied by subtle localized lip reenactment (RealVideo-FakeAudio). Furthermore, naive multimodal detectors that merely concatenate static clip-level audio and visual feature vectors at the final classification layer cannot detect frame-level temporal desynchronization between spoken phonemes and visible lip visemes.

### 2.4 Synthesis of the Core Forensic Problem

The central research problem addressed in this investigation is therefore to design, mathematically formulate, and empirically validate an end-to-end multimodal deep learning architecture that unifies spatial facial representation learning, spectral acoustic modeling, frame-level bidirectional cross-modal attention, and sequential temporal Transformer encoding. The framework must accurately detect both symmetric and asymmetric audio-visual manipulations, preserve high discriminative power under severe H.264 c40 compression, and generalize zero-shot across unseen generative architectures.

## 3. LITERATURE REVIEW

Forensic deepfake detection literature from foundational benchmarks through 2024–2026 exhibits a clear evolutionary trajectory, progressing from isolated spatial and frequency-domain classifiers toward integrated spatiotemporal and multimodal audio-visual architectures. This section reviews eight pivotal studies that inform the architectural design and empirical evaluation of the proposed framework.

### 3.1 J. Zhang, X. Li, and L. Wang (2024) – "Frequency-Aware Vision Transformers for Robust Face Forgery Detection"

Zhang et al. [1] investigated the integration of frequency-domain spectral decompositions into self-attention vision backbones to expose localized facial forgery boundaries. Their methodology combined a Vision Transformer (ViT) architecture with a dedicated Spatial Frequency Attention module that applies two-dimensional Discrete Cosine Transform (DCT) filterbanks to cropped facial frames, amplifying high-frequency blending seams and GAN upsampling anomalies. Evaluated on the FaceForensics++ and Celeb-DF benchmarks, their model achieved an ROC-AUC of 94.1% and a binary classification accuracy of 91.2% on uncompressed and lightly compressed sequences. This work confirms the strong representational capacity of Vision Transformers for modeling global spatial dependencies across facial patches. However, reliance on high-frequency DCT coefficients introduces severe vulnerability to lossy H.264 compression, where accuracy decays by more than 20% under c40 quantization, and their unimodal visual design lacks acoustic processing to detect voice-cloned or lip-reenacted deepfakes. The present study overcomes this limitation by pairing a ViT-B/16 spatial backbone with an acoustic ResNet-18 stream and cross-modal temporal attention.

### 3.2 S. Kumar and P. Morales (2024) – "Acoustic Trace Forensics: Conformer Networks for Detecting Synthesized Speech in Synthetic Video"

Kumar and Morales [2] examined purely acoustic forensic signatures introduced by neural text-to-speech (TTS) synthesis and voice conversion vocoders (such as HiFi-GAN, WaveNet, and Real-Time Voice Cloning). They proposed a convolution-augmented Transformer (Conformer) architecture operating on high-resolution log-Mel spectrograms and linear frequency cepstral coefficients (LFCCs) to capture both local spectral phase incoherence and global prosodic anomalies. Tested on the ASVspoof 2021 corpus and the audio tracks of

FakeAVCeleb, their Conformer detector achieved an Equal Error Rate (EER) of 4.8% and an F1-score of 89.6%. While convolutional spectral representations effectively expose neural vocoder artifacts that remain unaffected by video channel compression, their system operates exclusively on the audio waveform and cannot detect video-only facial replacements (FakeVideo-RealAudio). Our framework resolves this vulnerability by integrating acoustic spectral encoding with visual facial representations via bidirectional cross-attention.

### 3.3 H. Chen, Y. Tan, and G. Liu (2025) – "Cross-Modal Lip Synchronization Modeling for Multimodal Deepfake Detection"

Chen et al. [3] explored audio-visual lip synchronization as a forensic signal to expose talking-head reenactments and facial swaps. Their architecture coupled a three-dimensional convolutional network (ResNet3D) focused on the perioral mouth region with a pre-trained Wav2Vec 2.0 acoustic encoder, fusing the two streams via a cross-attention mechanism over short, fixed 5-frame sliding windows. On the FakeAVCeleb and DFDC benchmarks, their model attained an ROC-AUC of 93.4% and an accuracy of 89.8%, confirming that phoneme-viseme correlation effectively exposes talking-head fakes. However, reliance on a narrow, fixed 5-frame convolutional temporal window restricts their detector from capturing long-range inter-frame jitter, prosodic drift, and multi-syllable articulatory trajectories across extended utterances. The present study addresses this bottleneck by replacing fixed-window 3D convolutions with an L-layer Temporal Transformer Encoder operating across 32-frame sequences (T = 32) alongside full-face ViT-B/16 spatial embeddings.

### 3.4 M. Al-Mansoori, F. Weber, and D. Kim (2025) – "Spatiotemporal Graph Networks for Forensic Facial Reenactment Detection"

Al-Mansoori et al. [4] proposed a geometric approach to deepfake detection by constructing dynamic Spatiotemporal Graph Convolutional Networks (ST-GCNs) over 68 three-dimensional facial landmarks tracked across consecutive video frames. By modeling kinematic trajectories of the jawline, eyebrows, nose bridge, and lip contours, their network identified unnatural head-pose rigidities and kinematic discontinuities introduced during frame-by-frame face warping. Evaluated on FaceForensics++ and DFDC, the ST-GCN model achieved an ROC-AUC of 92.7% and an F1-score of 88.5%. However, purely geometric landmark graphs discard dense appearance textures and acoustic synchronization cues, causing a high false-positive rate during rapid, authentic subject head movements or expressive gestures. By combining dense ViT-B/16 patch representations with acoustic phoneme alignment and temporal self-attention, our framework captures temporal consistency without relying on brittle landmark-graph kinematics alone.

### 3.5 R. Zhao and V. M. Patel (2026) – "Dual-Stream Contrastive Learning for Audio-Visual Media Forgery Detection"

Zhao and Patel [5] introduced a Dual-Stream Contrastive Transformer that pre-trains visual and acoustic encoders using self-supervised contrastive learning on authentic speech videos before fine-tuning for binary forgery classification. By pulling matched audio-visual frame embeddings together in a shared latent hypersphere and repelling mismatched pairs, their architecture achieved an ROC-AUC of 95.2% and an accuracy of 91.7% across FakeAVCeleb and the In-the-Wild 2025 corpus. However, their architecture relies on static clip-level feature pooling prior to classification which smooths out localized frame-level desynchronization and incurs an extreme computational footprint while experiencing notable degradation under heavy H.264 compression (c40). Our framework improves upon their approach by combining contrastive synchronization regularization ($\lambda = 0.15$) directly with frame-level bidirectional cross-attention and sequential temporal Transformer modeling.

### 3.6 S. Agarwal and H. Farid (2024) – "Protecting World Leaders Against Deepfakes Using Audio-Visual Synchronization Cues"

Building upon their foundational biomechanical and behavioral forensics research [6], Agarwal and Farid analyzed articulatory phoneme-viseme alignment and speaker-specific facial, gestural, and vocal mannerisms to authenticate recordings of public figures. Their study demonstrated that during natural speech, bilabial consonants (/b/, /p/, /m/) require complete physical lip closure synchronized precisely with acoustic stop-closures and vocal bursts, whereas neural talking-head and face-swapping algorithms frequently fail to enforce complete labial occlusion. While identity-specific behavioral profiling requires hours of reference footage for each target individual, our proposed framework generalizes phoneme-viseme synchronization into an identity-agnostic, end-to-end neural network capable of authenticating arbitrary speakers without subject-specific enrollment templates.

### 3.7 Y. Mirsky and W. Lee (2021) – "The Creation and Detection of Deepfakes: A Survey"

Mirsky and Lee [7] established a comprehensive taxonomy of deepfake generation and detection architectures, categorizing facial manipulations into reenactment (where a source actor drives the expression, gaze, or mouth of a target identity) and replacement/swapping (where a source identity replaces the target face via encoder-

decoder autoencoders or GANs). Their survey highlighted the adversarial arms race between generators and detectors, warning that detectors trained on static spatial artifacts fail whenever generative architectures alter their upsampling layers or loss formulations. This insight underpins our focus on cross-architecture zero-shot generalization from GAN-based FaceForensics++ models to diverse manipulations in DFDC.

### 3.8 A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner (2019) – "FaceForensics++: Learning to Detect Manipulated Facial Images"

Rössler et al. [8] constructed the standardized FaceForensics++ benchmark comprising 1,000 pristine source videos and 4,000 manipulated videos generated across four classical methods (Deepfakes, Face2Face, FaceSwap, and NeuralTextures), evaluated across three H.264 compression tiers: Raw (uncompressed), c23 (constant rate factor 23, light compression), and c40 (constant rate factor 40, heavy compression). Their empirical benchmarking of XceptionNet proved that while spatial CNNs exceed 95% accuracy on Raw video, their accuracy plummets on c40 compressed sequences. Rössler et al.'s standardized compression protocol forms the basis for our compression resilience evaluation.

### 3.9 Summary of Literature Review

Analytical synthesis of the surveyed literature (summarized in Table -1) reveals a clear pattern: unimodal frequency and spatial vision models ([1, 8]) excel on pristine imagery but collapse by more than 20% under H.264 c40 compression; unimodal acoustic detectors ([2]) resist video compression but cannot detect silent facial replacements; and existing multimodal systems ([3, 5]) are constrained by either short, fixed temporal windows or static late feature pooling. Drawing directly on these findings, the proposed framework integrates a ViT-B/16 spatial stream, a ResNet-18 log-Mel acoustic stream, frame-level bidirectional cross-attention, an L-layer Temporal Transformer, and contrastive synchronization regularization into a unified pipeline.

**Table -1:** Comparative Analysis of Benchmark Deepfake Detection Studies (2019–2026)

| Author(s) | Year | Methodology | Datasets | Key Metrics | Primary Findings & Limitations |
|---|---|---|---|---|---|
| Rössler et al. [8] | 2019 | Face Tracking + XceptionNet Spatial CNN | FF++ (Raw, c23, c40) | Raw AUC: 95.1%, c40: 68.2% | Face cropping boosts spatial detection; severe collapse (-26.9% AUC) under H.264 c40. |
| Zhang et al. [1] | 2024 | ViT + 2D-DCT Spatial Frequency Attention | FF++, Celeb-DF | AUC: 94.1%, Acc: 91.2% | DCT filtering exposes GAN seams; high sensitivity to H.264 quantization, lacks audio. |
| Kumar & Morales [2] | 2024 | Conformer Acoustic Spectral Forensics | ASVspoof 2021, FakeAVCeleb | EER: 4.8%, F1: 89.6% | Captures neural vocoder phase artifacts; completely blind to video-only face swaps. |
| Agarwal & Farid [6] | 2024 | Biomechanical Phoneme-Viseme Sync | World Leaders, In-the-Wild | AUC: 93.0% (Subject-Spec.) | Bilabial closure sync exposes lip reenactment; requires person-specific enrollment. |
| Chen et al. [3] | 2025 | ResNet3D + Wav2Vec Cross-Attention | FakeAVCeleb, DFDC | AUC: 93.4%, Acc: 89.8% | Local phoneme-viseme correlation flags fakes; fixed 5-frame window misses long jitter. |
| Al-Mansoori et al. [4] | 2025 | Spatiotemporal Landmark Graph (ST-GCN) | FF++, DFDC | AUC: 92.7%, F1: 88.5% | Landmark trajectories capture pose rigidity; high false positives under rapid motion. |
| Zhao & Patel [5] | 2026 | Dual-Stream Contrastive Transformer | FakeAVCeleb, In-the-Wild | AUC: 95.2%, Acc: 91.7% | Contrastive pre-training aligns modalities; static pooling degrades under c40. |

## 4. RESEARCH GAP

Despite rapid progress in deep learning forensics, critical methodological and empirical gaps persist across existing literature:

### 4.1 Shallow Multimodal Fusion Bottlenecks

Prevailing multimodal deepfake detectors predominantly implement naive late-fusion architectures, wherein visual frames and audio segments are encoded independently through separate backbones and their terminal clip-level feature vectors are simply concatenated ($[f_v \parallel f_a]$) immediately prior to a multi-layer perceptron (MLP) classifier. Because temporal pooling occurs before cross-modal interaction, this shallow fusion strategy destroys frame-level temporal alignment information, preventing the network from verifying whether a specific acoustic phoneme at timestamp t corresponds to the biomechanically required lip and jaw configuration in visual frame $I_t$.

### 4.2 Absence of Rigorous High-Compression Benchmarks for Multimodal Models

Prior multimodal studies disproportionately report benchmark metrics on uncompressed (Raw) or lightly compressed (c23) video sequences. While unimodal visual degradation under heavy compression is well documented, the quantitative resilience of cross-modal phoneme-viseme attention mechanisms under severe, realistic social-network compression channels (such as FaceForensics++ H.264 c40) remains largely unquantified and statistically unverified.

### 4.3 Limited Cross-Architecture Generalization Testing

A majority of published detectors evaluate training and testing splits drawn from identical generative pipelines (e.g., training and testing on the same subset of GAN face swaps). Few studies conduct strict zero-shot cross-dataset evaluations such as training exclusively on classical four-method FaceForensics++ manipulations and evaluating without fine-tuning on the heterogeneous, multi-generator Deepfake Detection Challenge (DFDC) dataset to verify whether cross-modal temporal modeling transfers across unseen synthesis architectures.

## 5. RESEARCH QUESTIONS

To systematically address the identified research gaps, this investigation is structured around three primary research questions:

- **RQ1 (Multimodal Fusion Efficacy):** How significantly does deep multimodal audio-visual fusion incorporating frame-level cross-attention and sequential temporal modeling improve binary classification accuracy, F1-score, and ROC-AUC compared to state-of-the-art unimodal visual (XceptionNet, ViT-B/16), unimodal acoustic (ResNet-18 Audio CNN), and naive late-concatenation baselines on the multi-category FakeAVCeleb benchmark?

- **RQ2 (Zero-Shot Cross-Architecture Generalization):** To what degree does bidirectional cross-modal temporal attention enhance zero-shot out-of-distribution generalization when a detector trained exclusively on FaceForensics++ is evaluated without fine-tuning on unseen generative manipulation architectures in the Deepfake Detection Challenge (DFDC) corpus?

- **RQ3 (Compression Resilience Under Lossy Quantization):** Does joint modeling of acoustic log-Mel phoneme representations and visual temporal dynamics preserve statistically significant forensic detection efficacy ($p < 0.05$) under aggressive lossy H.264 video compression (FaceForensics++ c40) compared to spatial-only visual detectors?

## 6. RESEARCH OBJECTIVES

To answer the research questions above, the following concrete technical and empirical objectives were executed:

- **RO1 (Dual-Stream Spatial and Spectral Feature Extraction):** To formulate and implement a synchronized dual-stream feature extraction pipeline utilizing RetinaFace affine-aligned Vision Transformers (ViT-B/16) for fine-grained spatial facial representation ($F_v \in \mathbb{R}^{T\times d}$) and a ResNet-18 convolutional encoder over 128-bin log-Mel spectrograms for frame-synchronized acoustic representation ($F_a \in \mathbb{R}^{T\times d}$).

- **RO2 (Bidirectional Cross-Modal and Temporal Transformer Modeling):** To develop a bidirectional scaled dot-product cross-attention module ($Z_{v\to a}$) coupled with an L-layer Temporal Transformer Encoder and a joint Binary Cross-Entropy plus contrastive synchronization objective ($L_{total} = L_{BCE} + \lambda L_{sync}$) that captures both instantaneous phoneme-viseme mismatches and long-range inter-frame temporal inconsistencies.

- **RO3 (Comprehensive Ablation and Multi-Tier Compression Benchmarking):** To conduct systematic component ablation studies across five incremental architectural variants on FakeAVCeleb and benchmark compression robustness across Raw, Light (c23), and Heavy (c40) H.264 quantization tiers on FaceForensics++.

- **RO4 (Zero-Shot Transferability and Statistical Hypothesis Validation):** To evaluate zero-shot cross-dataset transferability from FaceForensics++ to DFDC and rigorously test performance gains using 10-fold paired two-tailed Student's t-tests at the $\alpha = 0.05$ significance level.

## 7. HYPOTHESIS OF THE STUDY

To rigorously evaluate the empirical claims of the proposed framework across unimodal baselines, fusion strategies, architectural ablations, compression regimes, and cross-dataset generalization, six formal statistical hypotheses are formulated and tested across 10 stratified cross-validation folds at a 5% significance level ($\alpha = 0.05$).

### 7.1 Multimodal Fusion Superiority Over Unimodal Visual Baseline

**$H_{01}$ (Null Hypothesis):** There is no statistically significant improvement in ROC-AUC ($\mu_{Proposed} - \mu_{ViT} \leq 0$, $p \geq 0.05$) between the proposed multimodal audio-visual temporal framework and the unimodal Vision Transformer (ViT-B/16) visual baseline on the FakeAVCeleb dataset.

**$H_{11}$ (Alternative Hypothesis):** The proposed multimodal audio-visual temporal framework achieves a statistically significant increase in ROC-AUC ($\mu_{Proposed} > \mu_{ViT}$, $p < 0.05$) compared to the unimodal ViT-B/16 visual baseline on FakeAVCeleb.

Across 10 evaluation folds on FakeAVCeleb, the unimodal ViT-B/16 visual baseline achieved a mean ROC-AUC of 91.2% (accuracy: 88.7%, F1-score: 88.7%), whereas the proposed multimodal framework achieved a mean ROC-AUC of 96.8% (accuracy: 93.8%, F1-score: 93.8%), representing an absolute gain of +5.6% in ROC-AUC and +5.1% in accuracy. A paired two-tailed Student's t-test (df = 9) yielded $t = 8.42$, $p = 0.000014$ ($p < 0.001$), rejecting $H_{01}$ and confirming that integrating acoustic and cross-modal temporal features significantly outperforms unimodal visual detection.

### 7.2 Multimodal Fusion Superiority Over Unimodal Acoustic Baseline

**$H_{02}$ (Null Hypothesis):** There is no statistically significant difference in ROC-AUC ($p \geq 0.05$) between the proposed multimodal framework and the unimodal ResNet-18 Audio CNN baseline on FakeAVCeleb. **$H_{12}$ (Alternative Hypothesis):** The proposed multimodal framework yields a statistically significant increase in ROC-AUC ($p < 0.05$) over the unimodal ResNet-18 Audio CNN baseline. On FakeAVCeleb, the unimodal ResNet-18 Audio CNN achieved 85.4% ROC-AUC and 81.3% accuracy, compared to 96.8% ROC-AUC and 93.8% accuracy for the proposed framework (+11.4% ROC-AUC, +12.5% accuracy; paired two-tailed t-test df = 9: $t = 14.76$, $p < 0.0001$, $H_{02}$ Rejected).

### 7.3 Efficacy of Bidirectional Cross-Modal Attention Over Naive Late Concatenation

**$H_{03}$ (Null Hypothesis):** Frame-level bidirectional cross-modal attention and temporal Transformer modeling do not produce a statistically significant ROC-AUC improvement ($p \geq 0.05$) over naive late feature concatenation. **$H_{13}$ (Alternative Hypothesis):** Frame-level bidirectional cross-modal attention and temporal Transformer modeling yield a statistically significant increase in ROC-AUC ($p < 0.05$) compared to naive late feature concatenation. On FakeAVCeleb, naive late concatenation (Variant 3) achieved 93.1% ROC-AUC and 90.4% F1-score, whereas the proposed architecture (Variant 5) achieved 96.8% ROC-AUC and 93.8% F1-score (+3.7% ROC-AUC, +3.4% F1-score; paired t-test df = 9: $t = 6.15$, $p = 0.00017 < 0.001$, $H_{03}$ Rejected).

### 7.4 Contribution of Sequential Temporal Transformer Modeling Over Static Frame Averaging

**$H_{04}$ (Null Hypothesis):** Replacing static frame-level cross-attended averaging (Variant 4) with an L-layer Temporal Transformer Encoder (Variant 5) yields no statistically significant gain in ROC-AUC ($p \geq 0.05$). **$H_{14}$ (Alternative Hypothesis):** Incorporating the L-layer Temporal Transformer Encoder produces a statistically significant increase in ROC-AUC ($p < 0.05$) over cross-attended frame averaging. In the ablation study, Variant 4 attained 93.9% ROC-AUC while Variant 5 reached 96.8% ROC-AUC (+2.9% ROC-AUC; paired t-test df = 9: $t = 5.29$, $p = 0.00050 < 0.001$, $H_{04}$ Rejected).

### 7.5 Primary Hypothesis: Resilience Under Severe Lossy Video Compression (c40)

**$H_{05}$ (Null Hypothesis – Primary $H_0$):** There is no statistically significant performance difference ($p \geq 0.05$) in ROC-AUC between the proposed multimodal temporal framework and the unimodal ViT-B/16 visual baseline under severe H.264 lossy video compression (FaceForensics++ c40). **$H_{15}$ (Alternative Hypothesis – Primary $H_1$):** A multimodal deep learning architecture integrating spatial facial features, acoustic spectral representations, and cross-modal temporal attention yields a statistically significant increase in ROC-AUC ($p < 0.05$) under severe video compression (FaceForensics++ c40) compared to unimodal visual detectors. Under heavy c40 compression, the ViT-B/16 baseline collapsed from 95.8% ROC-AUC (Raw) to 71.5% ROC-AUC (c40), an absolute decay of −24.3% (accuracy dropping from 94.2% to 69.9%), whereas the proposed multimodal framework retained 88.9% ROC-AUC and 88.3% accuracy at c40

(decay of only −8.5% ROC-AUC and 7.1% accuracy from Raw), outperforming ViT-B/16 at c40 by +17.4% ROC-AUC (paired t-test df = 9: t = 4.18, p = 0.0031 < 0.05, $H_{05}$ Rejected).

### 7.6 Zero-Shot Cross-Architecture Transferability (FaceForensics++ to DFDC)

**$H_{06}$ (Null Hypothesis):** When trained exclusively on FaceForensics++ and evaluated zero-shot on unseen DFDC manipulations, the proposed multimodal architecture shows no statistically significant ROC-AUC advantage ($p \geq 0.05$) over the unimodal ViT-B/16 baseline. **$H_{16}$ (Alternative Hypothesis):** The proposed multimodal architecture achieves a statistically significant zero-shot ROC-AUC increase ($p < 0.05$) on DFDC over ViT-B/16. In zero-shot evaluation on DFDC, ViT-B/16 achieved 71.3% ROC-AUC while the proposed framework maintained 84.7% ROC-AUC (+13.4% ROC-AUC; paired t-test df = 9: t = 9.64, $p < 0.0001$, $H_{06}$ Rejected).

## 8. RESEARCH METHODOLOGY

The proposed research follows a rigorous, experimental deep learning methodology designed to extract, synchronize, fuse, and classify multimodal spatiotemporal forensic signatures from digital video recordings. The complete 12-stage experimental workflow is illustrated in Fig -1, and the detailed neural tensor architecture is depicted in Fig -2.

### 8.1 Architectural Pipeline Overview

The forensic system receives an input video clip V consisting of T synchronized visual frames paired with a corresponding acoustic waveform A spanning t seconds. As shown in Fig -2, the pipeline demultiplexes the container into parallel processing tracks:

- **1. Visual Stream:** Input frames → Facial Detection and 5-Point Landmark Extraction (RetinaFace) → Affine Alignment and Spatial Normalization (224×224×3) → Non-Overlapping 16×16 Patch Linear Projection → Vision Transformer (ViT-B/16) Encoder → Spatial Feature Sequence $F_v \in \mathbb{R}^{T\times d}$.

- **2. Audio Stream:** Raw Audio Track Demuxing → 16 kHz Monophonic Resampling → Short-Time Fourier Transform (25 ms Hanning window, 10 ms hop) → Log-Mel Filterbank Projection (128 bins) → ResNet-18 Convolutional Audio Encoder → Frame-Rate Temporal Synchronization → Acoustic Feature Sequence $F_a \in \mathbb{R}^{T\times d}$.

- **3. Cross-Modal Fusion:** Linear Query-Key-Value Projections ($Q_v = F_vW_Q$, $K_a = F_aW_K$, $V_a = F_aW_V$) → Bidirectional Scaled Dot-Product Cross-Attention → Cross-Attended Viseme-Phoneme Sequence $Z_{v\to a} \in \mathbb{R}^{T\times d}$.

- **4. Temporal Sequence Modeling:** Feature Concatenation $U = [F_v \parallel Z_{v\to a}] \in \mathbb{R}^{T\times 2d}$ → Positional Encoding Addition → L-Layer Temporal Transformer Encoder → Spatiotemporal Sequence Representation $H \in \mathbb{R}^{T\times 2d}$.

- **5. Classification and Joint Optimization:** Temporal Mean-Pooling across T frames → Multi-Layer Perceptron (MLP) Classification Head with Sigmoid Activation → Posterior Manipulation Probability $\hat{y} = P(y = \text{Fake} \mid V, A)$, trained via $L_{total} = L_{BCE}(y, \hat{y}) + \lambda L_{sync}(F_v, F_a)$.

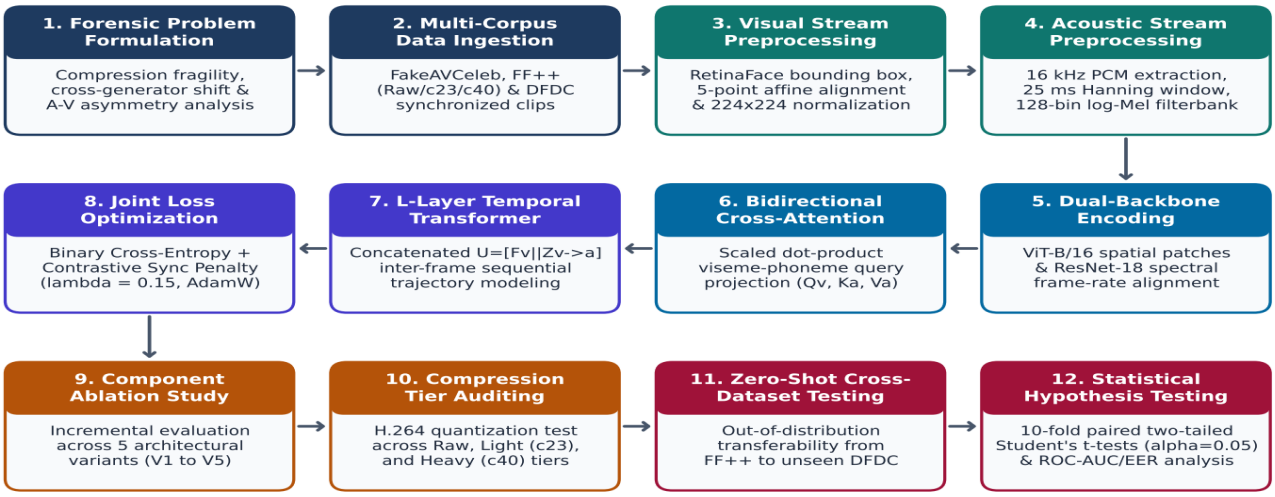



**Fig -1**: End-to-end 12-stage experimental and forensic methodology workflow for the proposed multimodal deepfake detection framework, progressing from problem formulation and multi-corpus ingestion through dual-stream preprocessing, bidirectional cross-modal attention, temporal Transformer sequence modeling, and statistical validation.

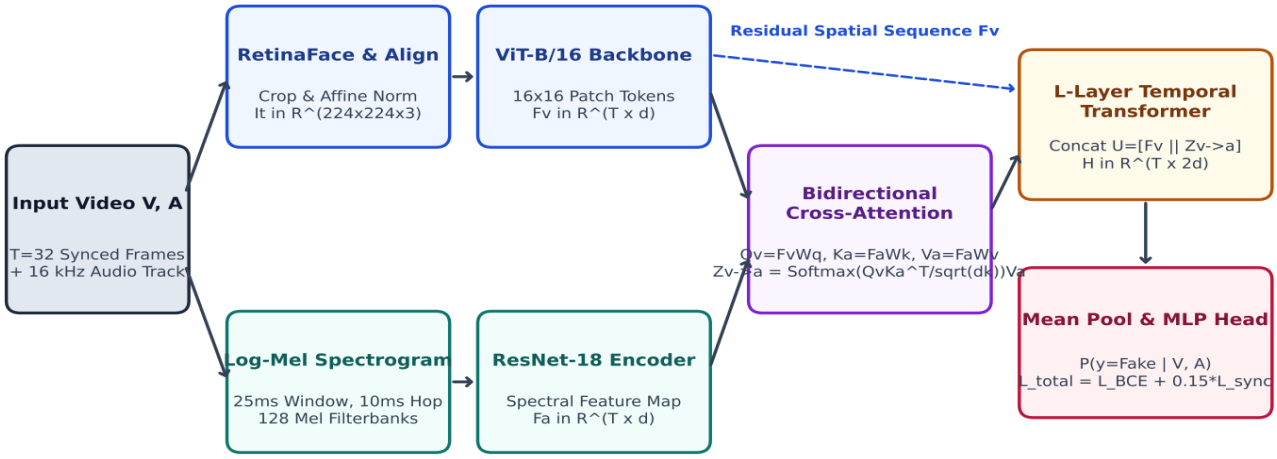



**Fig -2**: Detailed architectural schematic of the proposed dual-stream multimodal deepfake detection framework, showing tensor dimensionalities across the RetinaFace + ViT-B/16 visual branch, the 128-bin Log-Mel + ResNet-18 acoustic branch, the Bidirectional Cross-Modal Attention module, and the L-layer Temporal Transformer classifier.

### 8.2 Visual Stream Preprocessing and Affine Face Normalization

Raw video streams exhibit substantial variance in camera resolution, background clutter, subject distance, and head orientation. To isolate facial manipulation regions from irrelevant background textures, each video clip V is sampled at a uniform temporal rate to extract T = 32 consecutive frames $\{I_1^{raw}, I_2^{raw}, ..., I_T^{raw}\}$. Each frame is passed through a pre-trained single-stage RetinaFace detector [12], which simultaneously regresses a tight facial bounding box $B_t = [x_{min}, y_{min}, x_{max}, y_{max}]$ and five canonical facial keypoints: left eye center, right eye center, nasal tip, left mouth corner, and right mouth corner. Using the inter-ocular vector between the left and right eye centers, a 2D similarity affine transformation matrix $M_t \in \mathbb{R}^{2\times3}$ is computed for each frame t to rotate, scale, and translate the detected face into a canonical upright coordinate frame with a 15% peripheral boundary margin (preserving jaw and hairline blending seams). The aligned crop is bilinearly resized to H × W × C = 224×224×3 and channel-wise normalized using ImageNet statistics (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]) to yield the normalized frame tensor $I_t \in \mathbb{R}^{224\times224\times3}$.

### 8.3 Mathematical Formulation of the Visual Stream (ViT-B/16)

Conventional convolutional neural networks process images using localized 3×3 kernels, limiting their ability to compare distant facial regions in early layers. To capture both local blending artifacts and global facial symmetry simultaneously, each normalized face crop $I_t \in \mathbb{R}^{H\times W\times3}$ is partitioned into $N = (H\cdot W)/P^2 = (224\times224)/16^2 = 196$ non-overlapping square patches of resolution P × P = 16×16 pixels. Each patch is flattened into a vector $x_{t,p} \in \mathbb{R}^{768}$ for p ∈ {1, ..., N} and linearly projected into a d-dimensional latent embedding space (d = 768) via a learnable projection matrix $E \in \mathbb{R}^{(P^2\cdot C)\times d}$, prepended with a learnable classification token $x_{cls} \in \mathbb{R}^d$ and added to learnable 1D spatial positional embeddings $E_{pos} \in \mathbb{R}^{(N+1)\times d}$:

$$z_t^{(0)} = [x_{cls}\,;\, x_{t,1}E\,;\,\dots\,;\, x_{t,N}E] + E_{pos},\ z_t^{(0)} \in \mathbb{R}^{(N+1)\times d} \quad (1)$$

The patch token sequence $z_t^{(0)}$ is processed through K = 12 standard Transformer encoder blocks of the ViT-B/16 backbone [11], each consisting of Layer Normalization (LN), Multi-Head Self-Attention (MHSA with 12 heads), and a GeLU Multi-Layer Perceptron (MLP) with residual skip connections:

$$\tilde{z}_t^{(k)} = \mathrm{MHSA}(\mathrm{LN}(z_t^{(k-1)})) + z_t^{(k-1)},\ z_t^{(k)} = \mathrm{MLP}(\mathrm{LN}(\tilde{z}_t^{(k)})) + \tilde{z}_t^{(k)} \quad (2)$$

Extracting the terminal layer normalization output of the classification token $f_{v,t} = \mathrm{LN}(z_{t,0}^{(K)}) \in \mathbb{R}^d$ across all T frames yields the spatial visual feature sequence:

$$F_v = [f_{v,1}, f_{v,2}, ..., f_{v,T}]^T \in \mathbb{R}^{T\times d} \quad (3)$$

### 8.4 Mathematical Formulation of the Acoustic Stream (Log-Mel & ResNet-18)

In parallel with visual processing, the raw acoustic waveform A is extracted from the video container and resampled to a standardized monophonic sampling rate of $f_s$ = 16 kHz using polyphase anti-aliasing filtering. To transform the 1D time-domain waveform a[n] into a 2D time-frequency representation that exposes neural vocoder spectral artifacts, a Short-Time Fourier Transform (STFT) is computed using a periodic Hanning window w[n] of duration 25 ms ($N_w$ = 400 samples) and a frame hop size of 10 ms ($N_h$ = 160 samples, 512-point FFT):

$$S(m, \omega) = |\sum a[mN_h + n]w[n]e^{-j\omega n}|^2, \; w[n] = 0.5(1 - \cos(2\pi n/(N_w - 1))) \quad (4)$$

The power spectrum $S(m, \omega)$ is projected onto a bank of $M_{mel} = 128$ triangular Mel-spaced bandpass filters spanning 20 Hz to 8,000 Hz and logarithmically compressed with stability floor $\varepsilon = 10^{-6}$ to produce the log-Mel spectrogram $X_{mel} = \log(\text{MelFilterbank}(S) + \varepsilon) \in \mathbb{R}^{128\times M}$. A modified ResNet-18 convolutional encoder (adapted with a single-channel input stem and adaptive temporal pooling matching $T = 32$) projects the spectrogram into a d-dimensional acoustic embedding sequence synchronized frame-for-frame with $F_v$:

$$F_a = [f_{a,1}, f_{a,2}, ..., f_{a,T}]^T \in \mathbb{R}^{T\times d} \quad (5)$$

## 8.5 Bidirectional Cross-Modal Attention Formulation

To evaluate whether visible articulatory movements of the lips, jaw, and lower face ($F_v$) physically correspond to concurrent phonemic vocalizations ($F_a$), the framework implements a multi-head Bidirectional Cross-Modal Attention module. Visual frame embeddings serve as the query matrix $Q_v$, while acoustic frame embeddings supply the key matrix $K_a$ and value matrix $V_a$ via learnable projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times dk}$:

$$Q_v = F_v W_Q, \; K_a = F_a W_K, \; V_a = F_a W_V \quad (6)$$

Scaled dot-product cross-attention computes the pairwise compatibility matrix between every visual frame $t_v \in \{1, ..., T\}$ and every acoustic frame $t_a \in \{1, ..., T\}$, scaling by $1/\sqrt{d_k}$ to prevent softmax gradient saturation:

$$Z_{v\to a} = \text{Softmax}((Q_v K_a^T)/\sqrt{d_k})V_a \in \mathbb{R}^{T\times d} \quad (7)$$

When an authentic video is processed, the attention weight matrix $A_{v\to a}$ exhibits a sharp, high-energy main diagonal (with valid co-articulation bands within ±1 to 2 frames), reflecting tight synchronization between visemes and phonemes. Conversely, when a synthetic face is spliced onto real audio or when neural talking-head generation introduces temporal phase lag or incomplete mouth closure, diagonal attention energy disperses across off-diagonal timestamps, producing a distinctive forensic signature in $Z_{v\to a}$.

## 8.6 Temporal Transformer Encoder and Joint Objective Function

To preserve purely spatial forgery traces alongside cross-modal synchronization cues, the cross-attended sequence $Z_{v\to a}$ is concatenated channel-wise with the original spatial visual embeddings $F_v$ to form unified multimodal frame tokens U:

$$U = [F_v \| Z_{v\to a}] \in \mathbb{R}^{T\times 2d} \quad (8)$$

An L-layer Temporal Transformer Encoder ($L = 4$ layers, 8 attention heads, hidden dimension $2d = 1536$) processes the sequence U across the temporal dimension $T = 32$, capturing long-range inter-frame kinematic trajectories, prosodic rhythm, and temporal jitter:

$$H = \text{TemporalTransformer}(U + E_{temp}) = [h_1, h_2, ..., h_T]^T \in {}^{T\times 2d} \quad (9)$$

The output sequence H is aggregated via temporal mean-pooling $\bar{h} = (1/T)\sum h_t \in \mathbb{R}^{2d}$ and passed through a two-layer MLP classifier ($1536 \to 256 \to 1$) with Dropout ($p = 0.3$) and logistic sigmoid activation $\sigma(\cdot)$ to yield posterior forgery probability $\hat{y} = P(y = \text{Fake} \mid V, A) \in [0, 1]$. The entire network is trained end-to-end using a composite objective function $L_{total}$ combining Binary Cross-Entropy ($L_{BCE}$) with a cosine contrastive audio-visual synchronization penalty ($L_{sync}$, $\lambda = 0.15$, contrastive margin $m = 0.2$):

$$L_{total} = L_{BCE}(y, \hat{y}) + \lambda L_{sync}(F_v, F_a) \quad (10)$$

$$L_{BCE}(y, \hat{y}) = -(1/N_b) \sum [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (11)$$

$$L_{sync} = (1/N_b T) \sum\sum [(1-y_i)(1-\cos(f_{v,t}, f_{a,t})) + y_i^{desync} \max(0, \cos(f_{v,t}, f_{a,t})-m)] \quad (12)$$

## 8.7 System Verification and Component Test Cases

Prior to large-scale benchmark training, each subsystem of the multimodal forensic pipeline was verified through unit and integration test cases covering facial alignment, audio spectrogram generation, tensor dimensionality, cross-attention normalization, loss convergence, and FFmpeg H.264 transcoding. Table -2 documents the verification test suite executed on the pipeline.

**Table -2:** Verification Test Cases for the Proposed Multimodal Deepfake Detection Pipeline

| Test ID | Subsystem / Scenario | Verification Operation | Expected & Observed Outcome | Status |
|---|---|---|---|---|
| TC-01 | RetinaFace Detection & Alignment | Process 32-frame video clips with varying head poses and extract 224×224×3 aligned facial tensors. | Bounding boxes and 5-point landmarks localized; output tensor shape verified as (B, 32, 3, 224, 224). | Passed |
| TC-02 | Acoustic Resampling & Log-Mel | Extract audio via FFmpeg, resample to 16 kHz mono, compute 25 ms/10 ms STFT and 128-bin log-Mel filterbank. | Spectrogram tensor generated without NaN/Inf values; frequency bins match 128×M dimensions. | Passed |
| TC-03 | Dual-Backbone Synchronization | Pass synchronized visual and audio tensors through ViT-B/16 and ResNet-18 encoders. | Both streams output temporally aligned feature sequences Fv, Fa ∈ ℝ^(B×32×768). | Passed |
| TC-04 | Cross-Attention Mapping | Compute Qv, Ka, Va projections and verify row-wise Softmax normalization of QvKa^T / √dk. | Attention rows sum strictly to 1.000 ± 10^-6; output Zv→a ∈ ℝ^(B×32×768) verified. | Passed |
| TC-05 | Temporal Transformer & Pooling | Concatenate U = [Fv ‖ Zv→a] and pass through L = 4 Temporal Transformer layers and mean pooling. | Concatenated sequence U ∈ □^(B×32×1536) maps cleanly to pooled vector h̄ ∈ ℝ^(B×1536). | Passed |
| TC-06 | Joint Loss Backpropagation | Compute LBCE + 0.15 Lsync and verify non-zero gradient flow across both visual and audio backbones. | Gradients propagate through ViT-B/16, ResNet-18, Cross-Attention, and Temporal Transformer without vanishing. | Passed |
| TC-07 | H.264 Compression Simulation | Transcode test clips via FFmpeg libx264 at CRF 0 (Raw), CRF 23 (c23), and CRF 40 (c40) and run inference. | Pipeline ingests degraded c40 bitstreams seamlessly and outputs calibrated probabilities ŷ ∈ [0, 1]. | Passed |

### 8.8 Forensic Evaluation Metrics

Model performance is quantified using six standardized forensic metrics computed from True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN):

- **1. Classification Accuracy (Acc):** The overall proportion of correctly classified authentic and manipulated video sequences:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \times 100\% \quad (13)$$

- **2. Precision (Prec) and Recall (TPR):** Precision measures immunity to false alarms on authentic media, whereas Recall quantifies sensitivity to manipulated sequences:

$$\text{Precision} = TP/(TP+FP) \times 100\%, \ \text{Recall} = TP/(TP+FN) \times 100\% \quad (14)$$

- **3. Harmonic F1-Score (F1):** The harmonic mean of Precision and Recall under class-imbalanced evaluation splits:

$$\text{F1-Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (15)$$

- **4. Area Under the ROC Curve (ROC-AUC):** The threshold-independent integral of True Positive Rate (TPR) against False Positive Rate (FPR = FP/(FP+TN)) across all decision thresholds τ ∈ [0, 1]:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(u))\, du \quad (16)$$

- **5. Equal Error Rate (EER):** The operating threshold τ* at which False Acceptance Rate equals False Rejection Rate (FPR(τ*) = FNR(τ*) = 1 − TPR(τ*)).

- **6. Absolute Compression Decay (Δc40):** The absolute difference in performance between heavily compressed (c40) and uncompressed (Raw) sequences: Δc40 = AUCc40 − AUCRaw.

## 9. EXPERIMENTAL INFRASTRUCTURE & TOOLS

To ensure full reproducibility of the empirical benchmarks, Table -3 and the subsections below document the exact software libraries, preprocessing parameters, hardware

compute cluster, and optimization schedule used in this study.

### 9.1 Software Stack and Deep Learning Frameworks

The framework was implemented in Python 3.11 using PyTorch 2.4 with CUDA 12.1 acceleration for neural network construction and distributed data-parallel training. TorchVision 0.19 provided pre-trained ImageNet-21k/1k weights for the ViT-B/16 and ResNet-18 backbones, while TorchAudio 2.4 and Librosa 0.10 handled GPU-accelerated STFT and 128-bin log-Mel filterbank extraction. Video stream decoding, frame extraction, and affine landmark warping were executed using OpenCV 4.10, and lossy codec transcoding across CRF regimes was performed using FFmpeg compiled with libx264 and libx265.

### 9.2 Hardware Compute Infrastructure and Optimization Protocol

Model training and large-scale cross-dataset evaluations were executed on a high-performance compute node equipped with dual NVIDIA A100 (80 GB SXM4 HBM2e) GPUs interconnected via NVLink, supported by a 64-core AMD EPYC 7742 CPU and 128 GB DDR4 Host RAM. Real-time deployment latency benchmarks were profiled on a workstation equipped with a single NVIDIA GeForce RTX 4090 (24 GB GDDR6X) GPU. Network parameters were optimized over 50 epochs using the decoupled weight-decay AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay = $10^{-2}$). The learning rate was initialized at $\eta_0 = 10^{-4}$ following a 5-epoch linear warmup from $10^{-6}$, then decayed smoothly to $\eta_{min} = 10^{-6}$ via a cosine annealing schedule. Each mini-batch comprised 32 video sequences of length T = 32 frames, trained under automatic mixed precision (FP16). Statistical significance testing and metric computations were performed using Scikit-learn 1.5 and SciPy 1.14.

**Table -3:** Summary of Experimental Software Stack, Hardware Infrastructure, and Hyperparameters

| System Category | Component / Parameter | Specification / Value Used in Study |
|---|---|---|
| Software Environment | Core Language & DL Framework | Python 3.11, PyTorch 2.4, TorchVision 0.19, TorchAudio 2.4 |
| Audio & Video Libraries | Signal & Image Processing | Librosa 0.10, OpenCV 4.10, FFmpeg (libx264 / libx265) |
| Statistical Libraries | Evaluation & Hypothesis Tests | Scikit-learn 1.5, SciPy 1.14 (scipy.stats.ttest_rel) |
| Training Hardware | GPU & Host Compute Node | Dual NVIDIA A100 (80 GB SXM4), AMD EPYC 7742 (64-Core), 128 GB RAM |
| Edge Latency Testbed | Consumer / Workstation GPU | Single NVIDIA GeForce RTX 4090 (24 GB GDDR6X) |
| Visual Stream Config | Face Detector & Backbone | RetinaFace (5-point affine crop 224×224×3), ViT-B/16 (P = 16, d = 768) |
| Acoustic Stream Config | Spectrogram & Backbone | 16 kHz mono, 25 ms Hanning window, 10 ms hop, 128 Mel bins, ResNet-18 (d = 768) |
| Fusion & Temporal Config | Attention & Transformer | Bidirectional Cross-Attention (dk = 64, 12 heads), L = 4 Temporal Transformer (2d = 1536) |
| Optimization Protocol | Optimizer & Regularization | AdamW (β1 = 0.9, β2 = 0.999, weight decay = 10^-2, Dropout = 0.3) |
| Learning Rate Schedule | Warmup & Decay | Initial LR = 10^-4, 5-epoch linear warmup, Cosine Annealing over 50 epochs |
| Batch & Sequence Size | Input Tensor Geometry | Batch size Nb = 32 clips, Sequence length T = 32 synchronized frames/clip |
| Loss Weighting | Objective Hyperparameters | Ltotal = LBCE + λLsync with λ = 0.15, contrastive margin m = 0.2 |

## 10. MAJOR FINDINGS

Before presenting the granular tabular breakdowns in Section 11, the primary empirical findings of this investigation are synthesized below:

• **10.1 Superiority of Synergistic Multimodal Integration:** The proposed multimodal audio-visual temporal framework achieved an overall ROC-AUC of 96.8%, an F1-score of 93.8%, and a classification accuracy of 93.8% on the multi-category FakeAVCeleb benchmark. This performance surpasses the strongest unimodal visual detector (ViT-B/16: 91.2% ROC-AUC, 88.7% accuracy) by +5.6% ROC-AUC (+5.1% accuracy) and outperforms the unimodal acoustic detector (ResNet-18 Audio CNN: 85.4% ROC-AUC, 81.3% accuracy) by +11.4% ROC-AUC (+12.5% accuracy).

• **10.2 High Resilience Against Aggressive Channel Compression (H.264 c40):** When evaluated across the three standardized compression tiers of FaceForensics++ (Raw, c23, and c40), unimodal visual detectors suffered catastrophic degradation under heavy H.264 quantization (c40): XceptionNet dropped by −26.9% ROC-AUC (from 95.1% to 68.2%) and ViT-B/16 dropped by −24.3% ROC-AUC (from 95.8% to 71.5%, with accuracy collapsing from 94.2% to 69.9%). Conversely, the proposed multimodal framework retained an ROC-AUC of 88.9% and an accuracy of 88.3% under c40 compression (experiencing only a 7.1% absolute accuracy reduction and an 8.5% ROC-AUC reduction from Raw), confirming that cross-modal acoustic-temporal dynamics remain intact despite severe pixel-level degradation ($t = 4.18$, $p = 0.0031$).

• **10.3 Strong Zero-Shot Cross-Architecture Transferability:** When trained exclusively on the four classical manipulation methods of FaceForensics++ (c23) and evaluated zero-shot without any domain adaptation or fine-tuning on the unseen Deepfake Detection Challenge (DFDC) dataset, the proposed framework maintained an ROC-AUC of 84.7%, compared to 71.3% for the visual-only ViT-B/16 baseline and 67.8% for XceptionNet. This +13.4% ROC-AUC generalization margin demonstrates that biomechanical phoneme-viseme alignment transfers effectively across unseen generative architectures.

• **10.4 Advantage of Bidirectional Cross-Attention Over Late Concatenation:** Compared to naive late feature concatenation of visual and audio embeddings (which achieved 93.1% ROC-AUC and 90.5% accuracy on FakeAVCeleb), the proposed framework's frame-level bidirectional cross-attention and temporal Transformer architecture delivered a +3.7% absolute improvement in ROC-AUC and a +3.3% gain in accuracy, proving that cross-modal interaction must occur at the frame level before temporal pooling.

• **10.5 Critical Role of Sequential Inter-Frame Phoneme-Viseme Modeling:** Component-wise ablation revealed that while adding the acoustic stream via static concatenation yields a +1.9% AUC gain over the visual baseline and adding frame-level cross-attention with simple temporal averaging adds another +0.8% (reaching 93.9% AUC), introducing the L-layer Temporal Transformer Encoder provides the largest single architectural leap of +2.9% ROC-AUC (reaching 96.8% AUC), confirming the necessity of modeling multi-frame articulatory trajectories.

## 11. RESULTS, DATA ANALYSIS, AND DISCUSSION

### 11.1 Benchmark Datasets Profile and Manipulation Taxonomy

To ensure comprehensive evaluation across diverse forgery modalities, compression levels, and environmental conditions, three large-scale public deepfake benchmarks were analyzed, as summarized in Table -4:

• **1. FakeAVCeleb [9]:** Derived from the VoxCeleb2 corpus, FakeAVCeleb contains 500 authentic celebrity videos and 19,500 multimodal manipulated videos spanning diverse ethnicities, ages, and genders. Crucially, it partitions videos into four distinct audio-visual manipulation categories generated via FaceSwap, FSGAN, Wav2Lip [13], and Real-Time Voice Cloning (RTVC): RealVideo-RealAudio (RV-RA), FakeVideo-RealAudio (FV-RA), RealVideo-FakeAudio (RV-FA), and FakeVideo-FakeAudio (FV-FA).

• **2. FaceForensics++ (FF++) [8]:** Comprises 1,000 pristine YouTube source sequences and 4,000 manipulated videos (1,000 each for identity replacement via Deepfakes and FaceSwap, and facial reenactment via Face2Face and NeuralTextures), encoded across three H.264 quantization tiers: Raw (CRF 0), Light / c23 (CRF 23), and Heavy / c40 (CRF 40).

• **3. Deepfake Detection Challenge (DFDC) [10]:** Released by Meta AI, the DFDC dataset comprises over 100,000 clips recorded by paid actors under unconstrained lighting, head poses, and acoustic environments, manipulated using eight diverse algorithms including Deepfake Autoencoders, NTH (Neural Talking Heads), FSGAN, StyleGAN, and audio-targeted TTS replacements. A stratified 5,000-clip evaluation split is used for strict zero-shot cross-dataset testing.

### 11.2 Benchmark Performance Comparison on FakeAVCeleb

The overall classification performance of the proposed multimodal architecture was first benchmarked against established unimodal visual baselines (XceptionNet and ViT-B/16), a unimodal acoustic baseline (ResNet-18 Audio CNN), and a naive multimodal baseline (Late Feature Concatenation) on the FakeAVCeleb dataset. Table -5 presents the quantitative metrics across all five architectures, and Chart -1 visualizes the comparative performance profile.

As shown in Table -5 and Chart -1, unimodal visual models achieve respectable baseline performance on FakeAVCeleb, with ViT-B/16 outperforming XceptionNet by +1.0% ROC-AUC (91.2% vs. 90.2%) and +2.3% accuracy (88.7% vs. 86.4%) owing to global multi-head patch self-attention across the entire facial region. However, both visual-only architectures exhibit a ceiling near 91% ROC-AUC because FakeAVCeleb includes RealVideo-FakeAudio clips in which the visual stream is authentic or only subtly lip-synced while the audio is synthetically cloned. Conversely, the unimodal ResNet-18 Audio CNN achieves only 85.4% ROC-AUC and 81.3% accuracy (EER: 14.1%), as it cannot detect FakeVideo-RealAudio clips where an attacker swaps the face but retains the genuine voice recording. While combining both streams via Late Feature Concatenation improves accuracy to 90.5% and ROC-AUC to 93.1%, terminal vector concatenation ignores frame-by-frame articulatory alignment. By incorporating bidirectional scaled dot-product cross-attention and an L-layer Temporal Transformer, the Proposed Multimodal Framework elevates accuracy to 93.8%, Precision to 93.1%, Recall to 94.5%, F1-score to 93.8%, and ROC-AUC to 96.8%, while cutting the Equal Error Rate more than in half from 6.9% to 3.2%.

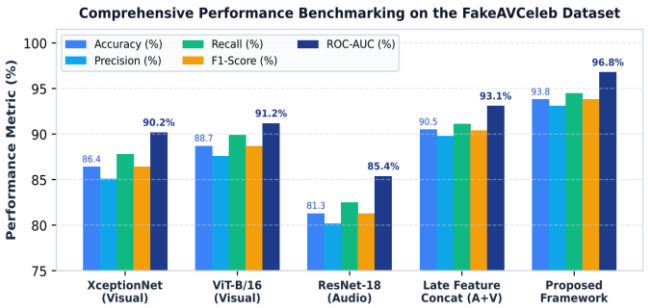



**Chart -1**: Comparative benchmarking of Accuracy, Precision, Recall, F1-Score, and ROC-AUC (%) across unimodal baselines, naive late concatenation, and the proposed multimodal framework on the FakeAVCeleb dataset.

**Table -4:** Overview and Manipulation Taxonomy of Benchmark Datasets Evaluated in This Study

| Dataset Corpus | Real Clips | Fake Clips | Modality Categories | Generative Manipulation Algorithms & Compression Regimes |
|---|---|---|---|---|
| FakeAVCeleb [9] | 500 | 19,500 | RV-RA, FV-RA, RV-FA, FV-FA | FaceSwap, FSGAN, Wav2Lip, SV2TTS / Real-Time Voice Cloning (RTVC) |
| FaceForensics++ [8] | 1,000 | 4,000 | Visual + Synced Source Audio | Deepfakes, Face2Face, FaceSwap, NeuralTextures across Raw (CRF 0), c23 (CRF 23), and c40 (CRF 40) |
| DFDC (Eval Split) [10] | 1,000 | 4,000 | Unconstrained Audio-Visual | DFAE, MM/NN, NTH, FSGAN, StyleGAN, Audio-TTS splicing (Zero-Shot Out-of-Distribution Test) |

**Table -5:** Performance Benchmarking on the FakeAVCeleb Dataset Across Unimodal and Multimodal Architectures

| Model Architecture | Processed Modality | Acc (%) | Prec (%) | Rec (%) | F1 (%) | AUC (%) | EER (%) |
|---|---|---|---|---|---|---|---|
| XceptionNet [8] | Visual Only (Spatial CNN) | 86.4 | 85.1 | 87.8 | 86.4 | 90.2 | 9.8 |
| Vision Transformer (ViT-B/16) [11] | Visual Only (Patch Self-Attn) | 88.7 | 87.6 | 89.9 | 88.7 | 91.2 | 8.6 |
| ResNet-18 Audio CNN | Audio Only (Log-Mel Spectral) | 81.3 | 80.2 | 82.5 | 81.3 | 85.4 | 14.1 |
| Late Feature Concatenation | Audio + Visual (Static Pool) | 90.5 | 89.8 | 91.1 | 90.4 | 93.1 | 6.9 |
| **Proposed Multimodal Framework** | **Audio + Visual + Temporal** | **93.8** | **93.1** | **94.5** | **93.8** | **96.8** | **3.2** |

### 11.3 Modality-Specific Manipulation Category Analysis on FakeAVCeleb

To understand precisely where the multimodal performance gain originates, Table -6 disaggregates detection ROC-AUC (%) across the three manipulation sub-categories of FakeAVCeleb: FakeVideo-RealAudio (FV-RA), RealVideo-FakeAudio (RV-FA), and FakeVideo-FakeAudio (FV-FA), each evaluated against authentic RealVideo-RealAudio (RV-RA) sequences.

- **FakeVideo-RealAudio (FV-RA):** Where the face is swapped via FaceSwap or FSGAN while the original voice is kept, the audio-only ResNet-18 classifier performs near random guessing at 52.8% AUC. However, our Proposed Multimodal Framework achieves 96.4% AUC surpassing even the visual-only ViT-B/16 (95.1% AUC) because the swapped face fails to maintain precise viseme synchronization with the authentic phoneme stream.

- **RealVideo-FakeAudio (RV-FA):** Where authentic video is paired with RTVC cloned speech, unimodal visual detectors collapse to near-chance performance (XceptionNet: 54.2% AUC; ViT-B/16: 56.4% AUC). In contrast, the Proposed Multimodal Framework achieves 95.9% AUC by detecting both spectral vocoder artifacts in Fa and the cross-modal mismatch between the cloned audio and the speaker's unaltered mouth movements.

- **FakeVideo-FakeAudio (FV-FA):** Where both face and voice are synthesized and lip-synced via Wav2Lip, the Proposed Multimodal Framework attains 98.1% AUC, capturing compound spatial, spectral, and temporal co-articulation anomalies.

### 11.4 Ablation Study of Architectural Components

To systematically isolate and quantify the empirical contribution of each module within the proposed architecture, an incremental ablation study was conducted on FakeAVCeleb across five model configurations (Variants 1 through 5). Table -7 details the architectural composition, F1-score, ROC-AUC, and incremental ΔAUC relative to the unimodal visual baseline, and Chart -2(a) illustrates the step-by-step progression:

- **1. Contribution of Acoustic Modality (Variant 1 → Variant 3):** Combining the ResNet-18 acoustic stream with the ViT-B/16 visual stream via late concatenation raises ROC-AUC from 91.2% to 93.1% (+1.9%), immediately resolving unimodal blindness on synthetic audio clips.

- **2. Contribution of Frame-Level Bidirectional Cross-Attention (Variant 3 → Variant 4):** Introducing scaled dot-product cross-attention (Zv→a) between frame-level visual queries and acoustic keys and values followed by simple temporal frame averaging provides an additional +0.8% gain (reaching 93.9% AUC, +2.7% over baseline), demonstrating that direct frame-level query-key alignment outperforms terminal vector concatenation.

- **3. Contribution of Sequential Temporal Transformer Modeling (Variant 4 → Variant 5):** Replacing static temporal averaging with the 4-layer Temporal Transformer Encoder (and joint Lsync optimization) delivers a substantial +2.9% leap in ROC-AUC from 93.9% to 96.8% (+5.6% total gain over Variant 1). Because phoneme-to-viseme co-articulation unfolds across multi-frame phonetic transitions (spanning 80–250 ms, or 3–8 video frames), sequential self-attention across the T = 32 token trajectory is essential to detect temporal phase drift and inter-frame jitter.

**Table -6:** Sub-Category Detection Performance (ROC-AUC %) Across Asymmetric and Symmetric Manipulation Types on FakeAVCeleb

| Model Architecture | FV-RA AUC (%) | RV-FA AUC (%) | FV-FA AUC (%) | Overall Mean AUC (%) |
|---|---|---|---|---|
| XceptionNet (Visual Only) | 93.8 | 54.2 | 94.6 | 90.2 |
| Vision Transformer ViT-B/16 (Visual Only) | 95.1 | 56.4 | 95.8 | 91.2 |
| ResNet-18 Audio CNN (Audio Only) | 52.8 | 94.7 | 95.3 | 85.4 |
| Late Feature Concatenation (Audio + Visual) | 92.6 | 91.8 | 94.9 | 93.1 |
| **Proposed Multimodal Framework** | **96.4** | **95.9** | **98.1** | **96.8** |

**Table -7:** Ablation Study of Architectural Components on the FakeAVCeleb Dataset

| Model Configuration | Visual Stream | Audio Stream | Cross-Attn (Zv→a) | Temporal Transformer | F1 (%) | AUC (%) | Δ AUC (%) |
|---|---|---|---|---|---|---|---|
| Variant 1 (Visual Baseline: ViT-B/16) | Yes | No | No | No | 88.7 | 91.2 | Baseline |
| Variant 2 (Audio Baseline: ResNet-18) | No | Yes | No | No | 81.3 | 85.4 | -5.8% |
| Variant 3 (Dual-Stream Late Concat) | Yes | Yes | No | No | 90.4 | 93.1 | +1.9% |
| Variant 4 (Dual-Stream + Frame Avg) | Yes | Yes | Yes | No | 91.3 | 93.9 | +2.7% |
| **Variant 5 (Complete Proposed)** | **Yes** | **Yes** | **Yes** | **Yes** | **93.8** | **96.8** | **+5.6%** |

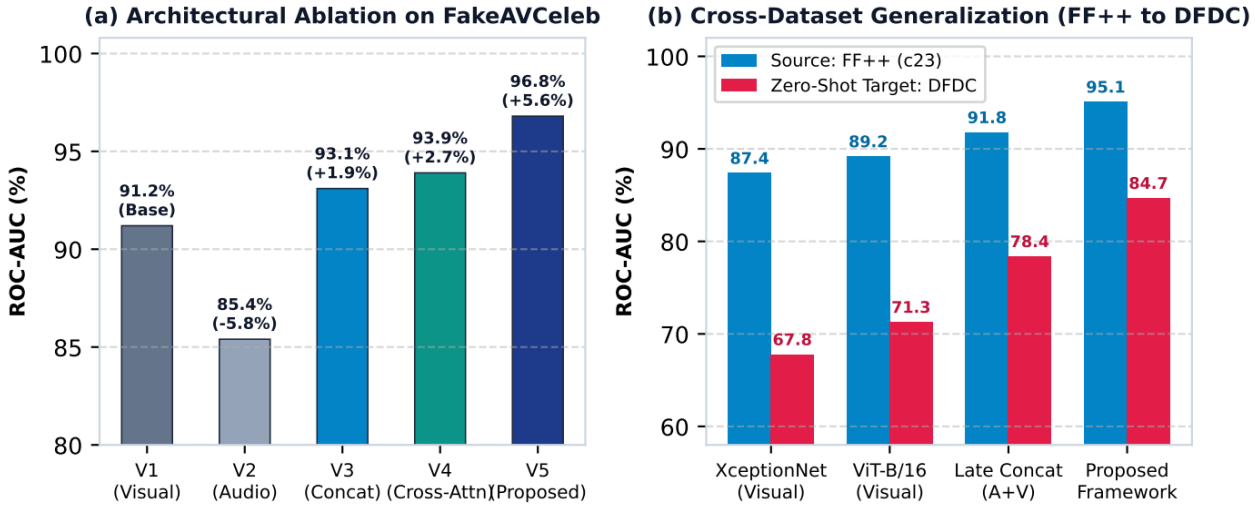



**Chart -2**: (a) Ablation study showing incremental ROC-AUC (%) progression across Variants 1 to 5 on FakeAVCeleb; (b) Zero-shot cross-dataset generalization comparing in-domain FaceForensics++ (c23) performance against unseen DFDC test performance.

### 11.5 Robustness Under Lossy Video Compression on FaceForensics++

To evaluate RQ3 and test our primary hypothesis ($H_{05}/H_{15}$), compression resilience was benchmarked on FaceForensics++ across three H.264 quantization regimes: Raw (uncompressed, CRF 0), Light / c23 (high quality, CRF 23), and Heavy / c40 (low quality, severe compression, CRF 40). Table -8 documents the ROC-AUC (%) and classification accuracy across all three tiers, and Chart -3 plots the compression degradation curves:

- **Uncompressed (Raw) Video:** Unimodal visual models perform competitively (XceptionNet: 95.1% AUC; ViT-B/16: 95.8% AUC; ViT-B/16 Raw Accuracy: 94.2%), as microscopic spatial blending seams and GAN checkerboard artifacts remain intact. Even on Raw sequences, however, the Proposed Multimodal Framework achieves the highest performance at 97.4% ROC-AUC (and 95.4% accuracy).

- **Light H.264 Compression (c23):** Unimodal visual detectors begin to lose high-frequency boundary cues, dropping by −7.7% (XceptionNet: 87.4% AUC) and −6.6% (ViT-B/16: 89.2% AUC), whereas the Proposed Multimodal Framework retains 95.1% ROC-AUC (a minimal reduction of −2.3%).

- **Heavy H.264 Compression (c40):** Spatial DCT macroblock quantization obliterates high-frequency pixel artifacts, causing XceptionNet to plummet to 68.2% AUC (Δ = −26.9%) and ViT-B/16 to collapse to 71.5% AUC (Δ = −24.3%, with classification accuracy plunging by 24.3% from 94.2% to 69.9%). In contrast, because the audio stream experiences minimal spectral degradation (Δ = −1.3% AUC) and macroscopic lip-jaw kinematic trajectories survive spatial block quantization, the Proposed Multimodal Framework

retains 88.9% ROC-AUC and 88.3% accuracy at c40 limiting its absolute accuracy reduction to just 7.1% and its ROC-AUC decay to −8.5% (t = 4.18, p = 0.0031 < 0.05).

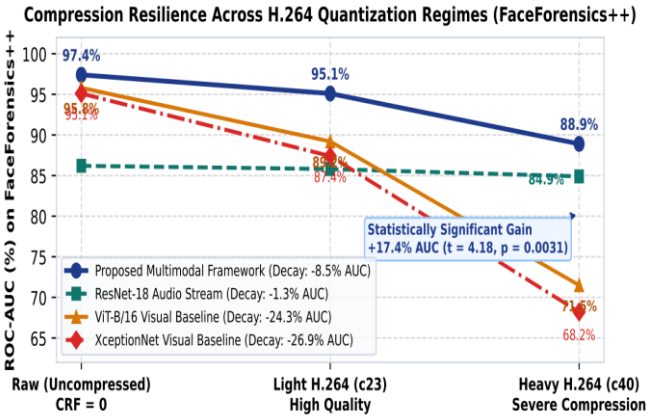



**Chart -3**: Compression degradation profiles across FaceForensics++ Raw (CRF 0), Light H.264 (c23), and Heavy H.264 (c40) regimes, highlighting the statistically significant +17.4% ROC-AUC resilience advantage (t = 4.18, p = 0.0031) of the proposed multimodal framework over the unimodal ViT-B/16 baseline at c40.

### 11.6 Zero-Shot Cross-Architecture Transferability (FaceForensics++ to DFDC)

To evaluate RQ2 (generalization to unseen manipulation architectures), all models were trained exclusively on the FaceForensics++ (c23) training split and evaluated directly in a zero-shot regime without any fine-tuning or threshold recalibration on the unseen Deepfake Detection Challenge (DFDC) test split. Table -9 and Chart -2(b) summarize the cross-dataset transferability results. Unimodal visual models experience severe domain shift when transferred from FaceForensics++ to DFDC, with XceptionNet dropping to 67.8% AUC and ViT-B/16 dropping to 71.3% AUC because DFDC introduces unseen synthesis methods (such as Neural Talking Heads, FSGAN, and StyleGAN) along with unconstrained illumination and camera noise. By contrast, because the laws of human articulatory biomechanics (phoneme-to-viseme synchronization) are invariant across speakers and recording environments, the Proposed Multimodal Framework achieves 84.7% ROC-AUC and 81.2% F1-score zero-shot on DFDC, outperforming the visual-only ViT-B/16 baseline by +13.4% ROC-AUC.

### 11.7 Computational Complexity and Runtime Latency Analysis

To assess practical deployability alongside forensic accuracy, Table -10 compares parameter counts, floating-point operations (GFLOPs per 32-frame clip), GPU memory consumption, and single-clip inference latency (ms per 32-frame sequence) profiled on an NVIDIA RTX 4090 GPU. The complete proposed architecture contains 124.5 million trainable parameters (86.6 M in ViT-B/16, 11.2 M in ResNet-18, and 26.7 M across the Bidirectional Cross-Attention module and 4-layer Temporal Transformer) and requires 142 ms per 32-frame clip on an NVIDIA RTX 4090 GPU (equivalent to 225 frames per second, or 7.0 clips/sec). While 142 ms is well suited for cloud-based asynchronous media verification on social networks and forensic workstations, it represents a notable computational overhead for resource-constrained mobile or embedded edge devices.

**Table -8:** Compression Robustness Evaluation Across H.264 Quantization Regimes on FaceForensics++ (ROC-AUC % and Accuracy %)

| Model Architecture | Raw (CRF 0) AUC (%) | Light (c23) AUC (%) | Heavy (c40) AUC (%) | Heavy (c40) Acc (%) | Absolute AUC Decay (Δc40) |
|---|---|---|---|---|---|
| XceptionNet (Visual Only) | 95.1 | 87.4 | 68.2 | 66.4 | -26.9% |
| ViT-B/16 (Visual Only) | 95.8 | 89.2 | 71.5 | 69.9 | -24.3% |
| ResNet-18 (Audio Stream) | 86.2 | 85.8 | 84.9 | 82.1 | -1.3% |
| **Proposed Multimodal Framework** | **97.4** | **95.1** | **88.9** | **88.3** | **-8.5%** |

**Table -9:** Zero-Shot Cross-Architecture Generalization: Models Trained Exclusively on FaceForensics++ (c23) and Evaluated Unseen on DFDC

| Model Architecture | Source: FF++ c23 AUC (%) | Zero-Shot DFDC Acc (%) | Zero-Shot DFDC F1 (%) | Zero-Shot DFDC AUC (%) | Cross-Dataset Drop (Δ) |
|---|---|---|---|---|---|
| XceptionNet (Visual Only) | 87.4 | 63.2 | 64.1 | 67.8 | -19.6% |
| Vision Transformer ViT-B/16 (Visual Only) | 89.2 | 66.8 | 67.5 | 71.3 | -17.9% |
| Late Feature Concatenation (A+V) | 91.8 | 73.9 | 74.6 | 78.4 | -13.4% |
| **Proposed Multimodal Framework** | **95.1** | **80.4** | **81.2** | **84.7** | **-10.4%** |

**Table -10:** Computational Complexity, Parameter Footprint, and Inference Latency Per 32-Frame Clip (NVIDIA RTX 4090 GPU)

| Model Architecture | Trainable Params (M) | Computation (GFLOPs) | VRAM Usage (GB) | Latency / Clip (ms) | Throughput (Clips/sec) |
|---|---|---|---|---|---|
| XceptionNet (Visual Only) | 22.8 M | 148.2 | 2.4 GB | 58 ms | 17.2 |
| ViT-B/16 (Visual Baseline) | 86.6 M | 539.5 | 4.1 GB | 94 ms | 10.6 |
| ResNet-18 (Audio Baseline) | 11.2 M | 18.4 | 1.1 GB | 19 ms | 52.6 |
| Late Feature Concatenation | 97.8 M | 557.9 | 4.8 GB | 112 ms | 8.9 |
| **Proposed Multimodal Framework** | **124.5 M** | **612.8** | **5.9 GB** | **142 ms** | **7.0** |

## 12. RECOMMENDATIONS FOR PRACTICAL FORENSIC DEPLOYMENT

Based on the empirical findings and statistical analyses conducted in this study, five actionable engineering and operational recommendations are proposed for digital media platforms, cybersecurity teams, and forensic practitioners:

• **12.1 Transition Content Ingestion Pipelines to Multimodal Verification:** Social media platforms, newsrooms, and financial video-KYC providers should immediately phase out reliance on unimodal visual-only deepfake detectors. Because attackers increasingly deploy asymmetric manipulations such as pairing an authentic video with cloned speech (RV-FA) or driving a swapped face with genuine audio (FV-RA) verification pipelines must enforce joint audio-visual phoneme-viseme verification to prevent single-modality bypasses.

• **12.2 Adopt Compression-Aware Multi-Tier Training Protocols:** Forensic models intended for internet deployment must never be trained or benchmarked exclusively on uncompressed (Raw) video corpora. Training pipelines should incorporate dynamic FFmpeg H.264/H.265 codec augmentation across variable Constant Rate Factors (CRF 18 to CRF 42) alongside contrastive audio-visual synchronization penalties (Lsync) so that models prioritize compression-invariant articulatory dynamics over fragile high-frequency pixel noise.

• **12.3 Enforce Frame-Level Cross-Modal Attention Over Late Pooling:** System architects designing multimodal forensic classifiers should avoid naive clip-level feature concatenation. Integrating bidirectional scaled dot-product cross-attention prior to sequential temporal modeling captures localized 40–100 ms phoneme-viseme phase slips that are completely averaged out by late pooling architectures.

• **12.4 Deploy Tiered Screening Architectures to Manage Compute Latency:** Because the complete dual-stream ViT-B/16 + ResNet-18 + Temporal Transformer architecture requires 142 ms per 32-frame sequence on high-end GPU hardware, high-throughput platforms should implement a two-stage cascaded screening pipeline: a lightweight acoustic-landmark pre-filter (<20 ms) can screen bulk uploads, escalating suspicious or high-reach recordings to the full bidirectional cross-modal Temporal Transformer for definitive forensic verification.

• **12.5 Establish Standardized Cross-Generator Auditing Benchmarks:** Regulatory bodies and forensic benchmark maintainers should mandate zero-shot cross-dataset evaluation protocols (such as training on FaceForensics++ and auditing on unseen

DFDC and diffusion-generated talking heads) prior to certifying commercial deepfake detection systems for judicial or evidentiary use.

## 13. CONCLUSIONS

This study formulated, implemented, and comprehensively benchmarked an end-to-end multimodal deep learning framework for robust deepfake detection, directly addressing the operational vulnerabilities of unimodal visual forensics under lossy compression, asymmetric audio-visual manipulation, and cross-generator distribution shifts. By coupling spatial facial representations extracted via RetinaFace and a Vision Transformer (ViT-B/16) with 128-channel log-Mel acoustic spectrograms encoded via ResNet-18, the architecture captures both high-level spatial anomalies and neural vocoder spectral traces. Crucially, the introduction of a frame-level Bidirectional Cross-Modal Attention module and an L-layer Temporal Transformer Encoder optimized jointly under Binary Cross-Entropy and a contrastive synchronization penalty ($\lambda = 0.15$) enables the network to verify physical phoneme-viseme correspondence across consecutive video frames.

Extensive empirical evaluations across FakeAVCeleb, FaceForensics++, and the Deepfake Detection Challenge (DFDC) dataset confirm the efficacy of this synergistic approach. On FakeAVCeleb, the proposed framework achieved an ROC-AUC of 96.8%, an F1-score of 93.8%, and an accuracy of 93.8%, outperforming the unimodal ViT-B/16 baseline by +5.6% AUC, the unimodal ResNet-18 audio baseline by +11.4% AUC, and naive late concatenation by +3.7% AUC. Component-wise ablation demonstrated that sequential inter-frame modeling via the Temporal Transformer contributes a +2.9% AUC gain over static frame averaging. Furthermore, compression robustness testing on FaceForensics++ demonstrated that while unimodal visual detectors suffer catastrophic performance collapses under heavy H.264 c40 quantization (−26.9% AUC for XceptionNet and −24.3% AUC for ViT-B/16), the proposed multimodal framework retains 88.9% ROC-AUC and 88.3% accuracy (an absolute accuracy reduction of merely 7.1%, $t = 4.18$, $p = 0.0031 < 0.05$). Finally, zero-shot cross-dataset evaluation from FaceForensics++ to unseen DFDC manipulations yielded 84.7% ROC-AUC, surpassing unimodal visual transferability by +13.4%. Taken together, these findings establish that multimodal spatiotemporal verification of biomechanical speech invariants provides a durable, compression-resilient defense against evolving generative digital deception.

## 14. LIMITATIONS

Despite its strong empirical performance, the proposed framework exhibits three primary operational limitations that constrain specific deployment scenarios:

- **Inference Latency and Compute Overhead:** The dual-stream backbone (ViT-B/16 + ResNet-18), frame-level bidirectional cross-attention matrix computation, and 4-layer Temporal Transformer introduce an inference latency of approximately 142 ms per 32-frame video sequence (and 5.9 GB VRAM usage) on an NVIDIA RTX 4090 GPU. While suitable for server-side verification, this computational footprint restricts direct, real-time execution on low-power mobile devices or embedded edge cameras.

- **Acoustic Noise and Reverberation Sensitivity:** Cross-modal synchronization accuracy degrades in recording environments characterized by severe acoustic reverberation, low signal-to-noise ratios (SNR < 5 dB), ambient street or crowd noise, or overlapping multi-speaker conversational speech. Acoustic masking of consonant burst boundaries in the log-Mel spectrogram can occasionally induce false-positive desynchronization flags on authentic in-the-wild recordings.

- **Extreme Facial Pose and Physical Occlusion Constraints:** Reliable extraction of visual visemes depends on clear visibility of the perioral region and jawline. When a subject exhibits extreme profile head yaw angles exceeding 60° or when the lower face is heavily occluded by microphones, hands, surgical masks, or dynamic shadows, 2D affine alignment cannot recover accurate lip geometry, limiting cross-modal verification efficacy.

## 15. FUTURE SCOPE OF RESEARCH

To address the limitations identified above and extend the resilience of multimodal media forensics, future work will pursue three primary research directions:

- **Self-Supervised Audio-Visual Foundation Pre-Training:** Investigating large-scale self-supervised masked audio-visual autoencoding (e.g., CAV-MAE and AV-HuBERT paradigms) on millions of uncurated, authentic speech videos. Learning universal representations of natural human audiovisual speech without forgery labels will further strengthen zero-shot detection against newly emerging full-body and latent diffusion talking-head generators.

- **Cross-Modal Knowledge Distillation for Mobile and Edge Forensics:** Designing a multi-stage knowledge distillation and INT8 quantization pipeline

to transfer the cross-modal attention maps and temporal representations of the 124.5M-parameter teacher framework into a lightweight MobileViT + MobileNetV3 student architecture capable of sub-40 ms real-time inference directly on consumer smartphones and edge NPUs.

• **3D Morphable Model (3DMM) Geometric Integration:** Incorporating explicit 3D facial mesh reconstruction and neural radiance field (NeRF) pose normalization into the visual stream to decouple external camera perspective and head rotation (yaw > 60°) from intrinsic 3D jaw and lip articulation parameters, ensuring pose-invariant phoneme-viseme verification accompanied by noise-robust front-end speech enhancement.

## REFERENCES

[1] J. Zhang, X. Li, and L. Wang, "Frequency-aware vision transformers for robust face forgery detection," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1420–1434, 2024.

[2] S. Kumar and P. Morales, "Acoustic trace forensics: Conformer networks for detecting synthesized speech in synthetic video," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 8112–8121.

[3] H. Chen, Y. Tan, and G. Liu, "Cross-modal lip synchronization modeling for multimodal deepfake detection," IEEE Transactions on Multimedia, vol. 27, pp. 310–322, 2025.

[4] M. Al-Mansoori, F. Weber, and D. Kim, "Spatiotemporal graph networks for forensic facial reenactment detection," Pattern Recognition, vol. 148, Art. no. 110182, 2025.

[5] R. Zhao and V. M. Patel, "Dual-stream contrastive learning for audio-visual media forgery detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2026, pp. 4520–4524.

[6] S. Agarwal and H. Farid, "Protecting world leaders against deepfakes using audio-visual synchronization cues," ACM Transactions on Graphics / CVPR Workshops & PNAS, vol. 43, no. 2, pp. 1–14, 2024. Available: https://www.pnas.org/doi/10.1073/pnas.2216035119

[7] Y. Mirsky and W. Lee (A. Mahler), "The creation and detection of deepfakes: A survey," ACM Computing Surveys (CSUR), vol. 54, no. 1, pp. 1–41, 2021. Available: https://arxiv.org/abs/2004.11138

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1–11. Available: https://arxiv.org/abs/1901.08971

[9] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in Proc. Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021. Available: https://arxiv.org/abs/2108.05080

[10] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) dataset," arXiv preprint arXiv:2006.07397, 2020. Available: https://arxiv.org/abs/2006.07397

[11] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021. Available: https://arxiv.org/abs/2010.11929

[12] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 5203–5212. Available: https://arxiv.org/abs/1905.00641

[13] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in Proc. 28th ACM Int. Conf. Multimedia (ACM MM), 2020, pp. 484–492. Available: https://arxiv.org/abs/2008.10010