

Article

A Multi-Agent Framework with Adaptive Orchestration for Explainable Multi-Modal Deepfake Detection

Saoirse Barry ^{1,*}, Susmitha Vekkot ² and Pancham Shukla ¹

¹ Department of Computing, Imperial College London, London SW7 2AZ, UK; panchamkumar.shukla@imperial.ac.uk

² Department of Electronics and Communication Engineering, Amrita School of Engineering Bengaluru, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India; v_susmitha@blr.amrita.edu

* Correspondence: saoirsebarry02@gmail.com

Abstract

Modern deepfakes leave forensic traces across audio, visual, and cross-modal channels, yet contemporary detectors are monolithic transformer-based architectures whose post hoc explanations struggle to isolate modality-specific artefacts. We propose a multi-agent system in which five specialised detection agents each target distinct forensic features: facial biometrics and image quality, ECAPA-TDNN speaker embeddings, cross-modal lip-sync, XceptionNet spatial analysis, and FreqNet frequency-domain analysis. Each agent is a self-contained detector with its own perception, decision and explanation. An adaptive orchestrator invokes supplementary agents when first-wave agents disagree; this escalation does not significantly improve in-distribution accuracy but adds forensic evidence, and its rate rises under domain shift. Each agent produces its own modality-appropriate explanation, and a report-generation agent synthesises these into a natural-language forensic summary. On the PolyGlottFake test set, a within-source measurement, the system reaches 99.91% accuracy at $\tau = 0.35$, set on the validation partition, with an AUC-ROC of 1.000. On the public MAVOS-DD benchmark the system detects forged speech, including unseen voice-conversion methods (AUC-ROC 0.79), while per-agent scores locate the gap in video-only forgeries (0.44); there, escalation significantly reduces errors, and one additional agent raises the overall AUC-ROC from 0.58 to 0.74, showing the framework's ability to ameliorate system weaknesses to unseen forgeries with additional agents.

Keywords: deepfake detection; multi-agent systems; explainable AI; multi-modal detection; adaptive orchestration; media forensics

Academic Editors: Antony Bryant, Paolo Bellavista, Kenji Suzuki, Horacio Saggion, Roberto Montemanni, Andreas Holzinger and Min Chen

Received: 5 July 2026

Revised: 26 September 2026

Accepted: 28 September 2026

Published: 2 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Deepfakes are synthetic audio-visual media generated using AI models, including Generative Adversarial Networks (GANs), diffusion models, and Variational Autoencoders [1]. These technologies enable the fabrication of hyper-realistic forgeries of humans, often without their prior consent [2]. As these models have become more powerful and accessible, human observers are no longer reliably able to discern between real and generated content [3], and the impacts of deepfakes are no longer speculative. In 2024, the engineering firm Arup was subject to a deepfake-based social engineering attack that caused a £20 million loss [4], and Deloitte's Center for Financial Services estimates that

deepfake-related incidents could enable financial fraud losses of up to US\$40 billion in the United States by 2027 [5]. The number of AI-generated videos online increased by 550% between 2019 and 2023 [6], and the applications now span financial fraud, political disinformation, and non-consensual intimate imagery [1]. Robust, trustworthy, and interpretable deepfake detection is therefore a matter of immediate societal importance [7].

Historical deepfake detection systems were primarily uni-modal, considering only the audio or visual stream in isolation. Such detectors can perform well in narrow experimental settings, but they often generalise poorly to different datasets [8] and are readily defeated by cross-modal forgeries that pair authentic video with synthetic audio, or vice versa. Multi-modal frameworks can outperform their uni-modal counterparts [8,9], partly because they leverage the strong correlation between visual lip movements and spoken audio [8]. However, the dominant multi-modal architectures, including hybrid CNN-transformers and Vision Transformers [10–12], operate as opaque black boxes [13,14]. This obscures the artefacts responsible for a classification and complicates both failure analysis and the calibration of user trust.

Explainable AI (XAI) techniques such as GradCAM, SHAP, LIME, and transformer attention mapping can highlight the features driving a prediction [13,15–17]. Explainability has benefits on both sides of the deployment interface. For users, it can help them recognise manipulation cues and calibrate their trust in consuming online content [18]. For developers, it supports debugging, the diagnosis of over-reliance on specific artefacts, and the construction of more adversarially robust systems [19]. However, XAI methods are typically applied retrospectively to fully trained models rather than integrated as core architectural components, which limits their diagnostic value because the underlying model was not designed with interpretability in mind. This mismatch is particularly noticeable for monolithic detectors that have to analyse many heterogeneous artefacts simultaneously, as the resulting explanations cannot cleanly attribute evidence across modalities (e.g., anomalous eye-blinking, inconsistent lip-sync, spectral voice-cloning artefacts).

Because deepfake forensic evidence is distributed across distinct channels, a similarly distributed architecture is well suited to the problem. This paper proposes a multi-agent framework in which specialised agents are each optimised for a specific forensic question, coordinated by an adaptive orchestrator, and each responsible for producing their own modality-appropriate explanation. The full pipeline and its behavioural flow are detailed in Section 3.2. The framework contributes:

- A five-agent detection architecture spanning facial biometrics, speaker embeddings, cross-modal lip-sync, spatial CNN analysis, and frequency-domain analysis. The system reaches 99.91% accuracy and a threshold-free AUC-ROC of 1.000 on the 2162-sample PolyGlottFake test set with equal agent weights at the validation-set threshold $\tau = 0.35$, using 57.1 M parameters (13.3 M trainable), which is less than half the size of a single ViT-based baseline. Unlike stacked CNN ensembles, the agents differ in modality and forensic role (Sections 2.3 and 4.3).
- Integrated multi-level explainability that combines GradCAM saliency, SHAP attributions, cross-modal attention, biometric overlays, and a language-model-synthesised natural-language forensic report; this is motivated by a systematic comparison of GradCAM, SHAP, and LIME.
- An adaptive disagreement-based orchestration mechanism that resolves around 96% of samples using only three agents while preserving near-ceiling accuracy, with the escalation rate itself acting as a distribution-shift indicator.
- An external evaluation on the public MAVOS-DD benchmark, reported by forgery family, generator and forged modality, together with a real-world stress test on a custom YouTube benchmark of BBC News and AI-avatar content, showing that the

per-agent scores and explainability outputs still provide useful diagnostics where accuracy degrades under distribution shift.

2. Related Work

This section organises prior work into five threads that the proposed framework draws upon or departs from: (i) uni-modal and multi-modal detection; (ii) transformer-based architectures; (iii) ensemble and multi-agent approaches; (iv) explainability methods; and (v) benchmark datasets. The concluding subsection (Section 2.6) sets out the research gap our framework targets.

2.1. Uni-Modal and Multi-Modal Detection

Uni-modal detection analyses a single stream, typically audio or visual, for manipulation artefacts [9]. CNNs dominate visual uni-modal detection, due to their ability to isolate forensic signatures [20] such as GAN fingerprints, inconsistent facial geometry, and localised textural signals from diffusion-based generators. Related work derives features from 3D facial geometry and illumination rather than image texture: Cheng et al. [21] mine temporal inconsistency between detected and 3D-reprojected facial landmarks, Tian et al. [22] apply a transformer encoder to illumination sequences from 3D face models, and Zhang et al. [23] fuse landmark normal vectors with illumination coefficients in a convolution-transformer hybrid, evaluated on the four FaceForensics++ C40 subsets, Celeb-DF, and DFDC. Physiological detectors additionally exploit inconsistencies in expected human behaviour, including abnormal eye-blinking, irregular ear and mouth movement, and anomalies in measured heart rate [24]. These methods are, however, vulnerable to signal inheritance, where the physiology of a driving video is inherited by the forgery; localised spatiotemporal phase-shift analysis has been proposed to expose regional inconsistencies [25]. Audio detectors similarly rely on CNNs over frequency-domain representations derived from the Discrete Cosine Transform or Short-time Fourier Transforms [26]. FreqNet, a uni-modal image detector operating on inverse and Fast Fourier Transforms, reaches 92.8% accuracy across seventeen generative models [27]; it is applied here to Mel spectrogram images (Section 3.3.2). As the generative landscape has shifted from GANs to diffusion and autoregressive transformers, the detection focus has correspondingly moved from periodic gridding artefacts to step-wise denoising inconsistencies and high-frequency spectral decay [28].

Multi-modal detection considers audio and visual streams jointly and has been shown to outperform uni-modal counterparts [9]. These methods exploit cross-modal inconsistencies arising from incompatibilities between streams [1,29]. Brief moments of lip desynchronisation can be exploited by architectures that embed visual and audio features into a joint latent space [17]; attention-based fusion then weights detected inconsistencies for strong classification [30]. The most recent contextual analyses employ Large Language Models (LLMs) and Vision-Language Models (VLMs) to evaluate contextual consistency and semantic coherence [31].

2.2. Transformer-Based Architectures

Hybrid CNN-transformer architectures combine the localised feature detection of CNNs with the long-range dependency modelling of transformers. Applied to uni-modal visual detection, this combination achieves 97.0% accuracy on FaceForensics++ using XceptionNet or EfficientNet feature extraction followed by a transformer encoder [11]. Vision Transformers (ViTs) are now widely used for feature extraction. GenConvIT couples a ConvNeXt CNN backbone with a Swin Transformer and attains 95.8% average accuracy across benchmarks [10], while Khormali and Yuan [12] reported 99.3% accuracy on the Celeb-DF dataset with an end-to-end vision transformer. LIPINC-V2 (April 2025)

uses a vision temporal transformer with multi-head cross-attention specifically for lip-sync inconsistency detection [32]. On the audio side, modern transformer-based approaches include Wav2Vec2-BERT and XLSR-MamBo; the latter uses linear-complexity State Space Model (SSM) embeddings for sequence encoding [26,33]. Frameworks such as HoliAntiSpoof [34] reformulate audio detection as a unified text-generation task using Audio Large Language Models, enabling joint reasoning over signal-level, paralinguistic, and semantic cues.

2.3. Ensemble and Multi-Agent Approaches

Ensembling combines outputs from multiple specialised models for more stable and reliable detection [35]. CNN ensembles feature prominently in the literature: combinations of InceptionV3, VGG16, and Xception reach 97.4% on the 140k Real-and-Fake Faces dataset [36], and ensembling has been repeatedly shown to outperform single-model components across diverse datasets [35]. DeepfakeStack is a representative example, using seven CNN base learners and a meta-learner to outperform every component architecture individually, reporting 99.65% accuracy with an AUROC of 1.0 [37]. Three properties distinguish such ensembles from the framework proposed here: their members are homogeneous visual CNNs sharing one input representation and failure modes; error correlation between members is not reported, so complementarity is an unexamined by-product of architecture; and every member runs on every input without producing a rationale.

Silva et al. [15] propose an explainable hierarchical ensemble of weakly supervised models; its members are all visual, and its explanations attach to the ensemble's decision rather than to each member's evidence (Section 3.6).

Multi-agent systems (MAS) use an autonomous nature and have distributed centres of control [38]. The canonical weak notion of agency requires four properties: autonomy, social ability, reactivity, and pro-activeness [39]. By dividing a global objective into sub-tasks dedicated to specialised agents, MAS can leverage complementary strengths [38,40]. Agents may use tools beyond pure language generation, and hierarchical orchestration, typically through a supervisory agent, coordinates system-wide control [40]. MAS are inherently fault-tolerant (a failed agent does not disable the system) and scalable, with additional agents extending capability [40]. Repeatable workflows with clear explainability traces benefit from a hierarchical architecture, while more decentralised approaches suit dynamic environments [41]. LLMs have introduced the ability for agents to communicate in natural language [42].

Multi-agent approaches to deepfake detection remain an emerging area. MAD-Sherlock uses multi-agent debate and consensus to identify manipulated images with clear natural-language rationales [42], and multi-agent deep-research frameworks for media verification combine reverse image search, metadata analysis, fact checking, and news processing under planner-agent orchestration [43]. The multi-modal structure of deepfake detection, with forensic signals spanning audio, visual, cross-modal, and contextual dimensions, is well suited to specialised agents. This offers a more future-proof and adaptive alternative to monolithic retraining. Within the MAS taxonomy, the system we propose is a hierarchical arrangement with a single orchestrator and heterogeneous specialist agents, a pattern that is under-explored in media forensics despite its use in adjacent domains such as medical decision support and financial analysis [44,45].

Against the four-property definition [39], the system has per-modality perception, conditional invocation (Section 3.5) and per-agent explanation (Section 3.6); its orchestrator is a deterministic rule-based policy (Section 3.5), so social ability and pro-activeness are absent, and the system can equally be read as a hierarchical ensemble of specialised classifiers under adaptive orchestration. Under either name, its distinction from stacked

ensembles is measured complementarity (maximum pairwise error correlation 0.307, Section 4.3).

2.4. Detection Explainability

There is a widening gap between the sophistication of detection architectures and the explainability methods applied to them. Pixel-level inconsistencies detected by CNNs are compatible with GradCAM, SHAP, and LIME, which highlight the input regions responsible for a prediction [13,15,46]. Slack et al. [47] show, however, that LIME and SHAP attributions can be adversarially manipulated, an argument for explainability as an architectural property rather than a post hoc overlay. Pixel-level explanations have been used successfully to identify blurred facial regions and anomalous harmonic frequencies [13,46,48].

Cross-modal detection operates at a higher level of abstraction. Some multi-modal architectures employ attention-weight mapping to highlight audio-visual segments contributing to a prediction. Self-attention scores from transformer-based architectures can be overlaid as heatmaps, and unlike LIME or SHAP, attention is intrinsic to the model rather than externally applied. Cross-modal attention extends this mapping across streams [49]. Attention weights show where a model was activated, rather than the forgery provenance [49]; Section 4.7 tests this by deletion before reading it as evidence.

A parallel line of work uses VLMs to generate natural-language rationales for detections [42,50]. Large Multi-Modal Models (LMMs) are increasingly applied to deepfake detection, leveraging vision-language alignment and Chain-of-Thought reasoning [51]. MAD-Sherlock orchestrates multi-agent debate for manipulated-image detection with natural-language justifications [42], and multi-agent deep-research systems combine tool use (reverse image search, metadata analysis, fact checking) under hierarchical orchestration to produce verification reports [43].

2.5. Deepfake Datasets

Deepfake datasets provide the foundation on which deep learning architectures are trained and benchmarked. Recent efforts have shifted toward multi-modal datasets and diverse manipulation pipelines (Table 1). DeepSpeak (2025) is a large-scale multi-modal benchmark of talking-head deepfakes built from recordings of consenting participants [52]. MAVOS-DD (2025) is a multilingual audio–video benchmark of YouTube-sourced speakers in eight languages forged by seven video generators, labelled by generator, language and forged modality [53]; it is the external benchmark of Section 4.11.

The PolyGlottFake dataset [54] was selected for this work because of the number and variety of manipulation techniques it includes (ten techniques across seven languages), which is better aligned with the heterogeneity of deepfakes encountered in the wild. Academic benchmarks are nonetheless known to diverge from real-world deepfakes [55], motivating the complementary evaluations on MAVOS-DD and on a YouTube set in Section 4.11.

Table 1. Overview of benchmark deepfake datasets since 2020.

Dataset	Year	Pristine	Fake	Total	# Methods	Type
DeeperForensics 1.0 [56]	2020	50,000	10,000	60,000	1	Visual
DFDC [57]	2020	23,654	104,500	128,154	8	Audio-visual
Celeb-DF [58]	2020	590	5639	6229	1	Visual
FFIW [59]	2021	10,000	10,000	20,000	1	Visual
KoDF [60]	2021	62,166	175,776	237,942	5	Visual
FakeAVCeleb [61]	2021	500	19,500	20,000	4	Audio-visual
DF-Platter [62]	2023	133,260	132,496	265,756	3	Visual

PolyGlottFake [54]	2024	766	14,472	15,238	10	Audio-visual
DeepSpeak [52]	2025	6226	6799	13,025	4	Audio-visual
DigiFakeAV [63]	2025	10,000	50,000	60,000	6	Audio-visual
MAVOS-DD [53]	2025	25,195	35,169	60,364	7	Audio-visual

2.6. Current Limitations and Framework Comparison

CNN-based detectors tend to overfit to specific manipulation artefacts and generalise poorly across datasets [27]. They are vulnerable to adversarial attacks that bypass fingerprint-specific features, for instance by processing the frequency domain or blurring “GAN fingerprints”. Specialised architectures can reach very high accuracy on their training distribution, but they typically degrade under domain shift [27]. A comparative overview of representative architectures is shown in Table 2, highlighting the trade-offs between specialised and generalisable designs.

Four gaps in the current literature motivate this framework. First, the monolithic design of contemporary state-of-the-art detectors does not match the distributed nature of deepfake forensic evidence. Second, explainability is almost always treated as a post hoc diagnostic rather than designed in from the start, which limits its fidelity. Third, static inference pipelines run the full model on every sample regardless of difficulty, missing the opportunity for adaptive, evidence-proportional analysis. Fourth, ensemble complementarity is assumed rather than measured, and members typically share one input modality (Section 2.3). The multi-agent framework introduced in Section 3 addresses all four.

Table 2. Representative deepfake detection architectures (2020 to 2026).

Architecture	Details	Results	Strengths and Weaknesses
HoliAntiSpoof [34]	Audio Large Language Model (ALLM) reformulating detection as a unified text-generation and reasoning task.	0.8% EER (SOTA)	Holistic analysis of paralinguistic and semantic cues; strong zero-shot and OOD generalisation through in-context learning.
XLSR-MamBo [33]	Hybrid Mamba-Attention (SSM) backbone combining selective state-space modelling with global attention.	98.5% Acc (OOD)	Linear O(L) complexity; captures both high-frequency signal artefacts and global spectral inconsistencies.
GenConViT [10]	Hybrid CNN-transformer with a Swin Transformer architecture and a ConvNeXt backbone.	95.8% avg	Strong generalisation; more computationally intensive than pure CNNs.
FreqNet [27]	Lightweight CNN classifier operating on the frequency domain.	92.8%	Lightweight model; image detector, applied here to Mel spectrogram images (Section 3.3.2).
XceptionNet [64,65]	Purpose-built CNN for deepfake detection.	95% (FF++)	High performance; better cross-dataset generalisation than ResNet or VGG16.
Hybrid CNN-LSTM [66]	CNN for spatial feature extraction and LSTM for temporal analysis.	94%	Strong spatial-temporal modelling; slow training and prone to vanishing gradients.
LIPINC-V2 [32]	Vision temporal transformer with multi-head cross-attention for lip-sync analysis.	0.99 AUC	Strong lip-sync detection; task-specific to lip manipulation.
Wav2Vec2-BERT [67]	Audio transformed into a 1D representation for sequence encoding.	91 to 99%	High performance; higher computational complexity than lightweight audio CNNs.
Hybrid CNN-Trans. [11]	EfficientNet-B4 or XceptionNet feature extraction with a transformer encoder.	97% avg	Higher generalisation than pure CNNs; transformer encoder can be computationally intensive.
End-to-end ViT [12]	Vision transformer for feature extraction and classification.	99.3% (Celeb-DF)	Best in class on visual-only datasets; no audio support; high computational cost.

3. Methodology

3.1. Dataset and Preprocessing

PolyGlottFake [54] was selected for its diversity of manipulation techniques and its multi-modal composition (Table 1). It contains 15,238 videos spanning seven languages (Arabic, English, Spanish, French, Japanese, Russian, Chinese), with a mean duration of 11.79 s at 1280 × 720 resolution. The dataset includes ten distinct multi-modal manipulation methods, the combinations of five voice-cloning and text-to-speech pipelines (XTTS, Vall-E-X, and Bark, Microsoft TTS (MicroTTS) or Tacotron each followed by FreeVC voice conversion) with two lip-sync methods (Wav2Lip with GAN enhancement and VideoRe-talking).

Train/validation/test splitting was performed per language with a fixed seed (42), applying a 70/15/15 ratio independently to the real and fake sets, so that the partition is reproducible (Table 3). Two partitions are used. The clip-level partition (Table 3, upper rows) splits each language 70/15/15 at the clip level; because PolyGlottFake derives about nineteen forgeries from each source video, 580 of the 709 test sources also appear in training (476 of the 697 forged test sources have their authentic clip there and 111 in validation), so it measures within-source generalisation. The source-disjoint partition (lower rows) keeps every clip of a source video, and every source in the same inferred facial-identity cluster, in one partition: face embeddings (InceptionResnetV1, VGGFace2) clustered by average linkage at a cosine distance of 0.40 group the 762 sources that have an authentic clip into 141 clusters and 69 singletons, and these units, with four sources that have only manipulated clips, are split 70/15/15 within language (seed 42) into 1182, 176 and 387 clips, with the test partition holding 83 authentic and 304 manipulated clips from 84 sources. Of the 3452 clips, two training clips (one authentic, one manipulated) have no preprocessed file and 1705 manipulated clips of training and validation sources are removed at random to balance the classes; Tables S1 and S2 give each partition's class, language and method composition. The clusters are inferred, not labelled. Stricter cuts only split them, so the partition stays disjoint at every cut from 0.20 to 0.40 (667 sources share a cluster at 0.20, 693 at 0.40). Adjacent source IDs mostly show the same person (the clustering merges 66% of adjacent-ID pairs against 1.7% of pairs 20 or more IDs apart), so an adjacent-ID neighbour in another partition marks a possible missed link: 48 of the 84 test sources have one in training or validation, or no embeddable authentic clip (Table S3), and Section 4.1 also reports the test read without them. Its agents are retrained under three seeds (Section 4.1) under a protocol fixed before any run. The training and validation sets of both partitions were downsampled for class balance; the test sets were left unbalanced.

Table 3. Full dataset split composition: clip-level partition (upper rows) and source-disjoint partition (lower rows, 1745 of the 3452 clips; Section 3.1 and Table S1).

Split	Real	Fake	Total
Training (downsampled)	526	526	1052
Validation (downsampled)	119	119	238
Testing (unbalanced)	118	2044	2162
Clip-level total	763	2689	3452
Source-disjoint training (downsampled)	591	591	1182
Source-disjoint validation (downsampled)	88	88	176
Source-disjoint testing (unbalanced)	83	304	387
Source-disjoint total	762	983	1745

The bold marks totals and the chosen configuration or operating point.

Visual preprocessing used MTCNN for face detection and cropping (confidence threshold 0.95). A frame stride of 10 balanced computational efficiency with temporal coverage. Extracted faces were resized to 299×299 pixels for XceptionNet and 224×224 for EfficientNet and ViT backbones. Audio preprocessing used librosa at 16 kHz, which is standard for speech-analysis tasks, with fixed 5 s zero-padding or cropping.

3.2. Pipeline Architecture

The pipeline comprises four sequential stages (Figure 1); agent is used in the scoped sense of Section 2.3. First, data pre-processing prepares the audio and visual streams for ingestion by the different models. This is followed by orchestrated multi-agent detection, with the results then aggregated to determine the classification. Finally, a report is generated to communicate the semantic context, system results and detection reasoning.

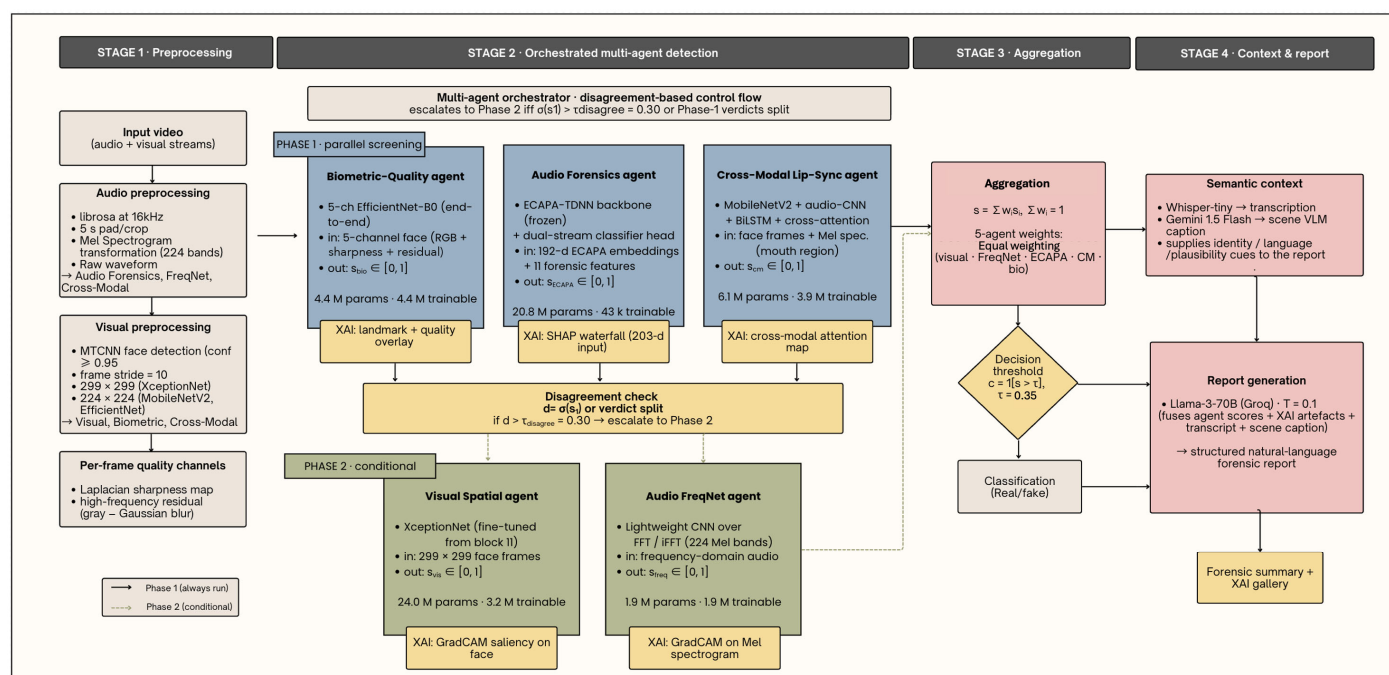


Figure 1. Architecture overview: five specialised detection agents (Phase 1: Biometric-Quality, Audio Forensics, Cross-Modal; Phase 2: Visual Spatial, Audio FreqNet) coordinated by the orchestration agent, each producing a modality-specific explainability output. Semantic-context and report-generation agents synthesise the per-agent evidence into a human-readable forensic summary.

Stage 1: Data pre-processing. The input video is decomposed into audio and visual streams. The visual stream undergoes MTCNN face detection (confidence ≥ 0.95) with frames sampled at stride 10. The audio stream is extracted with librosa at 16 kHz, standardised to 5 s, and transformed into Mel spectrograms (224 bands for FreqNet; the cross-modal agent computes its own 128-band representation, Section 3.3.3), MFCC features, and raw waveforms for embedding extraction. Per-frame image-quality maps are also computed: a per-pixel Laplacian sharpness map and a high-frequency residual map.

Stage 2: Orchestrated multi-agent detection. The orchestration agent coordinates deployment in two phases (Algorithm 1). Phase 1 (quick screening) dispatches three independent agents: the Biometric-Quality Agent, the Audio Forensics Agent (ECAPA-TDNN), and the Cross-Modal Lip-Sync Agent. Their inference calls can be executed concurrently as they share no state. The orchestrator computes an inter-agent disagreement metric; if it exceeds the disagreement threshold ($\tau_{disagree} = 0.3$), Phase 2 (deep analysis) is triggered, dispatching the Visual Spatial Agent (XceptionNet) and the Audio FreqNet

Agent. Each agent outputs a sigmoid probability and produces modality-specific explainability artefacts (Section 3.6).

Stage 3: Aggregation. The weighted sum of agent sigmoid outputs classifies the video: deepfake above the decision threshold, otherwise real.

Stage 4: Context and reporting. A Transcription Agent (Whisper tiny; OpenAI, San Francisco, CA, USA) and a Visual Context Agent (Gemini 1.5 Flash; Google, Mountain View, CA, USA) provide semantic context. Agent scores, explainability artefacts, transcripts, and scene descriptions are aggregated by a Report Generation Agent (Llama-3-70B; Meta, Menlo Park, CA, USA; via Groq, Mountain View, CA, USA; temperature 0.1) into a structured forensic report for human-readable output. The verdict is already fixed by the orchestrator (Equation (3), Algorithm 1), so the language model cannot alter it; detector-derived statements and contextual inferences occupy separate sections (Section 4.7).

Algorithm 1 Adaptive orchestration mechanism

Require: Video v ; weights w ; decision threshold τ ; disagreement threshold τ_{disagree}

Ensure: Classification $c \in \{\text{Real}, \text{Fake}\}$; explanation set E

```

1: Extract features: audio  $a$ , visual  $f$ , quality  $q$ 
2: Phase 1: Deploy agents in parallel
3:    $s_{\text{bio}} \leftarrow \text{BiometricAgent}(f, q)$ 
4:    $s_{\text{ecapa}} \leftarrow \text{EcapaAgent}(a)$ 
5:    $s_{\text{cm}} \leftarrow \text{CrossModalAgent}(a, f)$ 
6:    $s_1 \leftarrow [s_{\text{bio}}, s_{\text{ecapa}}, s_{\text{cm}}]$ 
7:    $d \leftarrow \sigma(s_1)$  ▷ Std of Phase 1 scores
8:   if  $\exists i, j: \mathbb{1}[s_i \geq 0.5] \neq \mathbb{1}[s_j \geq 0.5]$  then
9:      $d \leftarrow \max(d, \tau_{\text{disagree}})$  ▷ Force escalation on verdict split
10:  end if
11:  if  $d \geq \tau_{\text{disagree}}$  then
12:    Phase 2: Deploy supplementary agents
13:     $s_{\text{vis}} \leftarrow \text{XceptionAgent}(f)$ 
14:     $s_{\text{freq}} \leftarrow \text{FreqNetAgent}(a)$ 
15:     $s \leftarrow [s_1, s_{\text{vis}}, s_{\text{freq}}]$ ;  $w \leftarrow w_5$ 
16:  else
17:     $s \leftarrow s_1$ ;  $w \leftarrow w_3$ 
18:  end if
19:   $\bar{s} \leftarrow w \cdot s$  ▷ Weighted mean
20:   $c \leftarrow (\bar{s} \geq \tau) ? \text{Fake} : \text{Real}$ 
21:   $E \leftarrow \text{CollectExplanations}(s)$ 
22:  return  $c, E$ 

```

3.3. Individual Agent Development and Hyperparameter Selection

Each detection model was developed through systematic iterative experimentation prior to multi-agent integration. A consistent grid-search methodology was used for backbone selection, learning-rate tuning, and regularisation.

3.3.1. Visual Agent (XceptionNet)

The visual agent was selected through a comparative evaluation of CNN backbones. A baseline EfficientNetV2 (8,770,783 parameters; 1409 trainable), pretrained on ImageNet, achieved 92.08% accuracy on the English-subset test set. XceptionNet was subsequently evaluated for its documented effectiveness in deepfake detection through depthwise separable convolutions [65]. Initial training achieved 94.0% on the English-subset development evaluation, which improved to 98.6% through transfer learning, hyperparameter tuning, and face cut-out augmentation; the released agent scores 96.5% on the full test set

(Section 4.2). At inference, the agent averages its prediction over every sampled frame of the clip with horizontal-flip test-time augmentation, executed as one batched forward pass.

Transfer learning followed a two-stage schedule, first training the classifier head and then fine-tuning the backbone from block 11 onward, with hyperparameters selected by grid search on the validation set (Table S4, Text S1); the final model has 24.0 M parameters, of which 3.2 M are trainable.

3.3.2. Audio Agent (FreqNet)

Audio detection was developed through a systematic comparison of spectrogram representations and CNN backbones (Table S5), in which Mel spectrograms (97.6%) and MFCCs (97.4%) scored well above LFCC and CQT (86.8% and 84.2%), consistent with PolyGlottFake's voice-cloning artefacts lying in the Mel-frequency range. FreqNet [27] was adopted as the final audio agent (98% accuracy with FFT/iFFT features over 224 Mel bands). The released checkpoint was then fine-tuned on the training split alone under deployment-style waveform corruption and replaced only under the validation-only adoption rules of Section 3.5 (Text S1).

3.3.3. Cross-Modal Agent (CNN-BiLSTM with Cross-Attention)

The cross-modal agent captures lip-sync inconsistencies between audio and visual streams. The architecture uses MobileNetV2 (pretrained on ImageNet) for visual feature extraction with Global Average Pooling yielding a 1280-dimensional feature vector, and a two-block audio CNN over Mel spectrograms (128 bins $\times$ 313 time steps, Conv2d-ReLU-MaxPool stacked twice) feeding a bidirectional LSTM. A visual bidirectional LSTM models frame sequences in both temporal directions, and cross-attention (Equations (1) and (2)) is applied between audio and visual feature sequences. The model has 6.1 M parameters (3.9 M trainable). Training uses clip-consistent flip and rotation, frame dropout, SpecAugment-style masks, a weight decay of 0.05 and label smoothing 0.05. The released checkpoint was then fine-tuned on the training split alone under deployment-style corruption of both streams and adopted under the corrupted-validation rule of Section 3.5 (Text S1).

$$\text{attn_weights} = \text{softmax} \left(\frac{qK^T}{\sqrt{d_k}} \right) \quad (1)$$

$$\text{cross_modal_context} = \text{attn_weights} \cdot V \quad (2)$$

where q represents audio-clip features projected through learned weight matrices, and K , V are temporal visual feature representations of the mouth region. Cross-attention proved critical: 99% validation accuracy was achieved versus 92% with simple concatenation fusion.

3.3.4. Biometric-Quality Agent

The model is an EfficientNet-B0 whose first convolution is widened to five input channels: the RGB face crop, a tanh-compressed Laplacian sharpness map and a high-frequency residual (grayscale minus Gaussian blur). Training draws a random frame per clip per epoch under flip, rotation, colour-jitter, blur and erasing augmentation; a brief fine-tune is stopped on validation loss. Inference averages all frames with horizontal-flip test-time augmentation (99.2% validation accuracy). Landmark overlays are retained in the explainability outputs to provide geometric evidence, even though classification is driven by the five-channel CNN.

3.3.5. Audio Embedding Agent (ECAPA-TDNN)

The ECAPA-TDNN model produces a 192-dimensional speaker embedding, which is concatenated with 11 hand-crafted forensic features to form a 203-dimensional input representation. The hand-crafted features comprise five prosodic, three spectral, two temporal and one embedding-variance statistic (Text S1). A dual-stream classifier projects the ECAPA embedding to 64 dimensions and the forensic features to 32 dimensions; the two streams are concatenated and passed through a fusion head producing a sigmoid probability. The 192-dimensional ECAPA backbone is frozen during training; only the 43 k-parameter dual-stream classifier head is trained from scratch. SHAP analysis over this 203-dimensional input shows that the ECAPA embedding features are the dominant predictors, with spectral centroid and pitch statistics the most influential hand-crafted signals. These capture unnatural vocal variations characteristic of voice-cloning methods. The released checkpoint was then fine-tuned on the training split alone, each training clip embedded clean and under two corruptions, and adopted under the clean-validation rule of Section 3.5 (Text S1).

3.4. System Parameter Budget

Table 4 summarises total and trainable parameter counts for the five agents. The full ensemble totals 57.1 M parameters, of which 13.3 M are trainable. This is less than half the size of a single ViT-based deepfake baseline (≈ 130 to 180 M parameters), and the trainable count, which dominates training cost and over-fitting risk, is an order of magnitude smaller.

Table 4. Parameter counts per agent. Trainable counts reflect the frozen-backbone configuration used at inference (ECAPA-TDNN backbone frozen; XceptionNet fine-tuned from block 11; Biometric-Quality and FreqNet trained end-to-end; Cross-Modal MobileNetV2 backbone partially frozen).

Agent	Total Params	Trainable Params	Notes
Visual (XceptionNet)	24.0 M	3.2 M	Fine-tuned from block 11
Audio (FreqNet)	1.9 M	1.9 M	End-to-end training
Cross-Modal (MobileNetV2 + BiLSTM)	6.1 M	3.9 M	Backbone partially frozen
Biometric-Quality (5-ch EfficientNet-B0)	4.4 M	4.4 M	End-to-end training
Audio Forensics (ECAPA + 11 features)	20.8 M	0.04 M	ECAPA backbone frozen
System total	57.1 M	13.3 M	—

3.5. Adaptive Orchestration and Threshold Selection

The orchestration agent aggregates deployed agent scores using a weighted mean,

$$\bar{s} = \sum_{i=1}^N w_i \cdot s_i, \quad \sum_{i=1}^N w_i = 1 \quad (3)$$

where $N = 3$ in Phase 1 and $N = 5$ when Phase 2 is triggered (N being the number of agents). The weights are equal, $(w_{\text{visual}}, w_{\text{FreqNet}}, w_{\text{ECAPA}}, w_{\text{cross-modal}}, w_{\text{biometric}}) = (0.20, 0.20, 0.20, 0.20, 0.20)$, fixed a priori rather than fitted: the five agents reach comparable validation accuracy, so no agent is given greater weighting. In Phase 1 the three weights renormalise to one, so the first-wave verdict is the plain mean of the three scores (three-agent rows in Sections 4.1 and 4.4).

The inter-agent disagreement metric combines the standard deviation of agent scores with a verdict-disagreement indicator:

$$d = \begin{cases} \max(\sigma(s), \tau_{\text{disagree}}) & \text{if } \exists i, j: \mathbb{1}[s_i \geq 0.5] \neq \mathbb{1}[s_j \geq 0.5] \\ \sigma(s) & \text{otherwise} \end{cases} \quad (4)$$

where $\sigma(s)$ is the standard deviation of the Phase 1 agent scores and $\mathbb{1}[\cdot]$ is the indicator function applied to each agent's individual sigmoid verdict at the agent's own midpoint

of 0.5. If any two Phase 1 agents disagree on the verdict (one predicts fake, another predicts real), Phase 2 escalation is forced regardless of score spread; otherwise, escalation depends on the natural spread of scores. This ensures that categorical disagreement, which carries more forensic weight than a small numerical gap, always triggers the second wave.

The aggregate score is binarised at $\tau = 0.35$, set on the validation partition to minimise the false-negative rate: under equal weights the partition's minimum error is one, reached for every τ from 0.277 to 0.386, where the single error is a false alarm (a real scoring 0.414), and again above 0.414 up to 0.595, where it is a missed fake (scoring 0.386), with both errors in the narrow gap between; 0.35 lies inside the lower band (Section 4.5.1), 0.074 above the second-highest real and 0.036 below the lowest fake, so no validation fake is missed, while the conventional 0.5 would instead miss one validation fake. Agent verdicts in the escalation test of Algorithm 1 use each agent's own midpoint of 0.5, so escalation does not depend on τ . The disagreement threshold $\tau_{\text{disagree}} = 0.3$ was fixed during development as a compute-cost control (Section 4.5.2). Every reported number is re-derivable from the released per-sample scores; validation scores accompany the code. No parameter was tuned on the test set: the weights are fixed and the threshold is set on the validation partition. Two checkpoint-adoption rules govern the fine-tunes of Section 3.3, and both read the validation partition alone. Under rule 1, the clean-validation rule, a fine-tuned epoch replaces its released checkpoint when it improves clean validation log-loss and AUC-ROC. Under rule 2, the corrupted-validation rule, applied where clean validation loss does not improve, the epoch replaces the checkpoint only when it improves validation loss under the same corruption applied to the validation clips. Table 5 summarises the orchestration parameters.

Table 5. Orchestration hyperparameters.

Parameter	Value	Method
Aggregation	Weighted mean	Equation (3)
Agent weights w_i (5-agent)	(0.20, 0.20, 0.20, 0.20, 0.20)	Equal; fixed a priori (Section 3.5)
Decision threshold τ	0.35	Set on the validation partition to minimise missed fakes (Section 4.5.1)
Disagreement threshold τ_{disagree}	0.30	Fixed during development

3.6. Integrated Explainability Pipeline

Explainability is integrated as a core architectural component rather than applied post hoc. Table 6 maps each agent to its XAI method, output type, and the forensic question it addresses. The GradCAM formulation for CNN-based agents is:

$$L_{\text{GradCAM}} = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right), \quad \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (5)$$

Table 6. Agent-to-explainability mapping: each agent produces modality-specific explanations addressing distinct forensic questions.

Agent	XAI method	Output Type	Forensic Question
Visual (XceptionNet)	GradCAM	Saliency heatmap on face	Which facial regions show manipulation artefacts?
Audio (FreqNet)	GradCAM	Saliency on Mel spectrogram	Which frequency bands contain anomalous patterns?
Cross-Modal (Lip-Sync)	Cross-attention weights	Frame-level attention	At which temporal moments does audio-visual sync break?

Biometric-Quality	Landmark + quality overlay	Geometric and quality map	Are facial geometry and image quality authentic?
Audio Embedding (ECAPA-TDNN)	SHAP waterfall	Feature attribution plot	Which vocal profile features indicate synthetic speech?
Transcription (Whisper)	ASR	Text transcript	What is being said? (for semantic analysis)
Visual Context (Gemini)	VLM scene description	Natural language	Who is shown and is the context plausible?
Report Generation (Grog)	LLM synthesis	Structured NL report	What is the overall forensic assessment?

XAI Methods

GradCAM, SHAP, and LIME were systematically compared on both visual and audio baseline models. GradCAM and LIME showed partial overlap in highlighted regions for both modalities; SHAP provided more granular pixel-level attribution but was less interpretable for high-level artefacts. A key observation was that applying all three simultaneously produced noticeable discrepancies that could yield confusing explanations in deployment. GradCAM was selected as the primary CNN explainability method based on (i) computational efficiency (intrinsically computed gradients rather than perturbation-based methods); (ii) visual interpretability for face and spectrogram modalities; and (iii) maximal overlap with both LIME and SHAP outputs. LIME was excluded from final outputs due to the computational cost of pixel-perturbation calculations, which was incompatible with deployment requirements.

3.7. Baseline Models

We implemented and trained two published transformer-based models on PolyGlottFake as baselines: GenConViT [10] (≈ 130 M AE/140 M VAE parameters) and LIPINC-V2 [32] (4.9 M parameters). Two custom architectures were also developed: a Late-Fusion Vision Transformer (ViT with Wav2Vec2-BERT; 181 M parameters) and a Hybrid CNN-Transformer (EfficientNet-B0 + transformer encoder; 6.42 M parameters) using cross-modal attention between audio MFCC features and visual mouth regions. All benchmarks were augmented with GradCAM, SHAP, LIME, and attention mapping for explainability comparison.

We trained all five baselines under the per-language 70/15/15 split (seed 42; Section 3.1) and scored them on the same 2162-sample test partition as the proposed framework; Section 4.10 lists epoch budgets and parameter counts. Beyond those budgets the baseline training configurations were not retained, nor were per-sample predictions, so every baseline figure in this paper is a point estimate from our implementation, a single unpaired run that supports no inferential comparison (Section 5).

4. Results and Analysis

This section presents an end-to-end evaluation of the framework: benchmark comparison against state-of-the-art and custom transformer baselines (Section 4.1); individual agent characterisation through accuracy, AUC, score separation, and inter-agent correlation (Sections 4.2 and 4.3); systematic ablations and sensitivity analyses (Sections 4.4 and 4.5); training dynamics (Section 4.6); qualitative explainability analysis and case studies (Sections 4.7 and 4.8); an external evaluation on the public MAVOS-DD benchmark and on a YouTube set (Section 4.11); and a measured inference-cost comparison (Section 4.12).

4.1. Performance Comparison with Baseline Models

Table 7 reports full performance on the PolyGlottFake test set. All baselines were trained under the protocol of Section 3.7.

Table 7. Performance comparison on the PolyGlotFake test set (2162 samples: 118 real, 2044 fake). All multi-agent rows were evaluated at $\tau = 0.35$; baseline rows are single-run point estimates from our implementations (Section 3.7). The custom-designed baselines (Hybrid CNN-Transformer, Late-Fusion ViT) are reported in Section 4.10. The last row is the source-disjoint partition of Section 3.1 (387 clips: 83 real, 304 fake), with every agent retrained under three seeds.

Model	Acc.	Prec.	Rec.	F1	XAI
<i>Baselines (our implementations)</i>					
GenConViT AE	97.3	98.1	97.3	97.5	Post hoc
GenConViT VAE	95.3	97.4	95.3	95.9	Post hoc
LIPINC-V2	91.2	92.2	89.3	91.0	Post hoc
<i>Proposed multi-agent system</i>					
Multi-Agent (3-agent)	99.91	99.90	100.00	99.95	Integrated
Multi-Agent (5-agent)	99.91	99.90	100.00	99.95	Integrated
Multi-Agent (5-agent), source-disjoint, mean of three seeds	98.45	98.07	100.00	99.02	Integrated

The five-agent aggregate score completely separates the real and fake classes: AUC-ROC 1.000 and average precision of 1.000 (Figure 2), with the highest real aggregate being 0.413 and the lowest fake 0.480. At $\tau = 0.35$ the system classifies 2160 of 2162 test samples correctly (99.91%; two false positives, no missed fakes), a within-source measurement (Section 3.1).

On the source-disjoint partition (Table 3, lower rows), with every agent retrained under seeds 42, 43 and 44 and $\tau = 0.35$ held fixed (inside every run's zero-error validation band), the accuracy is 98.4% (standard deviation of 0.8 points; last row of Table 7), the AUC-ROC is 0.9999 to 1.0000 and the average precision is 1.0000. Without the 48 test sources flagged as possible identity overlaps (Section 3.1), the remaining 36 sources give 97.7% (standard deviation 1.2 points; 96.5%, 97.7% and 98.8%), still with no missed fakes and an AUC-ROC 0.9998 to 1.0000, against 99.1% on the flagged sources (Table S3), so residual overlap may flatter the headline by about one point. Every error is a false alarm on one of the 83 authentic clips (9, 6 and 3 per seed, that is 97.7%, 98.4% and 99.2%; specificity of 89.2% to 96.4% against 98.3% on the clip-level partition, and sensitivity of 100%). The visual agent, which reads the face itself, is the one most affected by faces from identity clusters absent from training (AUC-ROC of 0.951, down from 0.998) and accounts for the residual false alarms (removing it leaves one to five errors per seed); it is kept for the spatial evidence it renders, which escalation delivers to the analyst. Escalation fires on 8.8% to 11.6% of clips against 4.4%.

Our implementations of the baselines give point estimates of 97.3% (GenConViT AE), 91.2% (LIPINC-V2) and 98.1% (hybrid CNN-Transformer); as single unpaired runs they support no claim of superiority in either direction (Section 3.7). At $\tau = 0.35$ the three-agent Phase-1 configuration (the mean of ECAPA-TDNN, cross-modal and Biometric-Quality) makes 2 errors (99.91%), the same count as the deployed system; escalation changes which samples they are, not how many. The similar scores of the three-agent and five-agent configurations reflect that 95.6% of samples are not escalated to the five-agent system, as they fall below the disagreement threshold. Phase 2 does not significantly improve in-distribution accuracy: the trio and the five-agent system each make two clip-level errors and differ by at most 4 on the source-disjoint partition (trio 13, 4, 5 against 9, 6, 3; exact McNemar $p = 0.13, 0.69, 0.50$). In distribution, escalation is thus a compute-cost control that adds evidence (the saliency maps of Section 4.7) and a shift indicator (Section 4.9); whether it adds accuracy under shift is measured in Section 4.11, where the five agents make significantly fewer errors than the trio. Section 4.4 isolates each agent's individual contribution.

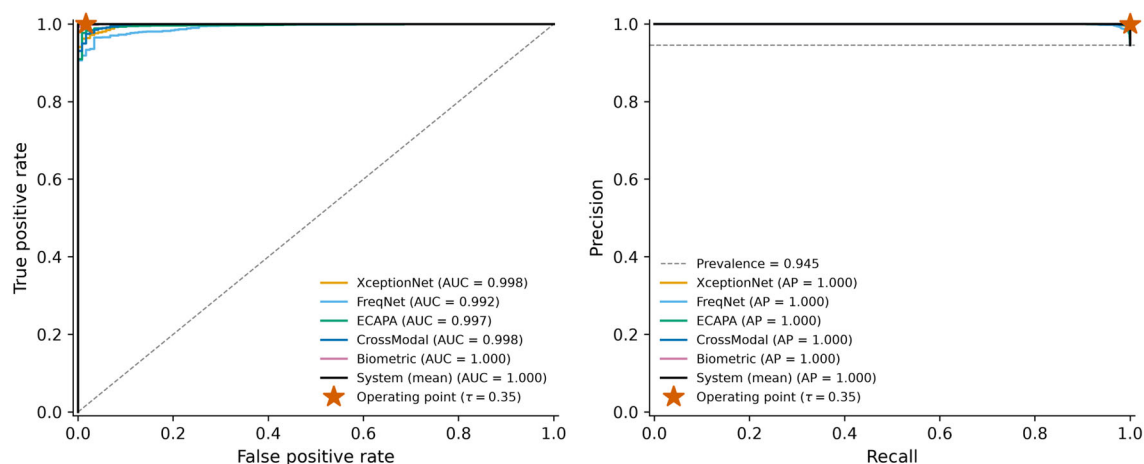


Figure 2. System-level ROC (left) and precision-recall (right) curves for the five-agent aggregate score on the PolyGlottFake test set, with per-agent curves overlaid. AUC-ROC = 1.000 and AP = 1.000; the operating point at $\tau = 0.35$ is marked.

4.2. Individual Agent Performance

Table 8 lists per-agent metrics: Biometric-Quality leads on accuracy (99.3%) and AUC (0.9999); Cross-Modal 98.0%, ECAPA-TDNN 97.8%, frame-averaged visual 96.5%; FreqNet is weakest (93.6%) yet highly precise (99.8%). The visual agent and FreqNet have the lowest accuracies but precision of 99.9% and 99.8%: they are almost always correct when they call a clip fake, and their lower recall is compensated by the other agents in the ensemble.

Table 8. Individual agent performance on the full test set.

Agent	AUC	Acc.	Prec.	Rec.	F1
Visual (XceptionNet)	0.9977	96.5	99.9	96.4	98.1
Audio (FreqNet)	0.9916	93.6	99.8	93.3	96.5
Audio (ECAPA-TDNN)	0.9966	97.8	99.9	97.8	98.8
Cross-Modal (Lip-Sync)	0.9980	98.0	99.8	98.1	98.9
Biometric-Quality	0.9999	99.3	99.9	99.3	99.6

The ROC curves (Figure 3) show all agents well above the random baseline, with four of the five approaching an AUC of 1.0. XceptionNet and FreqNet show the most gradual curves, consistent with their role as Phase 2 arbitrators rather than primary classifiers.

4.3. Score Distribution and Inter-Agent Analysis

Analysis of agent score distributions yields three design-relevant findings. First, Figure 4 reports the mean agent scores by content type and the score separation between deepfake and real predictions. Biometric-Quality exhibited the largest score separation (0.925), with the cross-modal (0.867) and ECAPA-TDNN (0.823) agents close behind. The visual agent showed the lowest separation (0.691), consistent with its lower standalone accuracy.

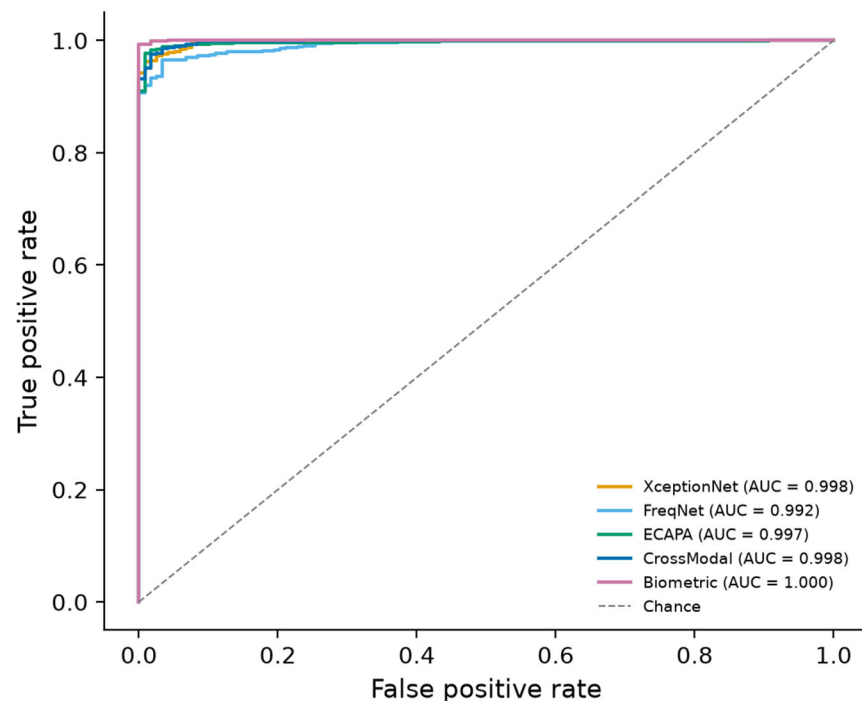


Figure 3. ROC curves for the five individual agents. Biometric-Quality (AUC = 0.9999), Cross-Modal (AUC = 0.9980) and Visual (AUC = 0.9977) exhibit near-ideal performance; FreqNet (AUC = 0.9916) shows the most gradual curve.

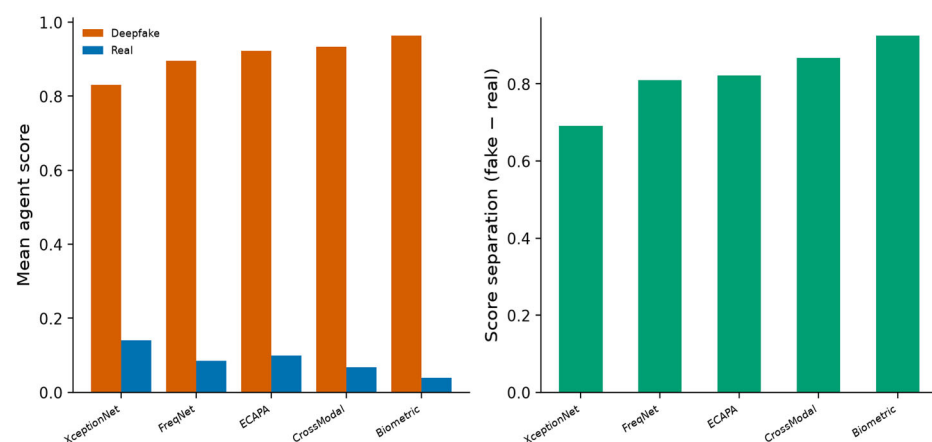


Figure 4. Agent score separation analysis. **(left)** Mean agent score by content type; **(right)** score difference (deepfake minus real) per agent.

Second, Figure 5 reports the mean prediction distance from the true label per agent. ECAPA-TDNN, Cross-Modal and Biometric-Quality had the tightest distances (0.079, 0.066 and 0.037), confirming their reliability as standalone classifiers. FreqNet (0.103) and the visual agent (0.167) showed the highest variability; they are less reliable individually but valuable as Phase 2 arbitrators for high-uncertainty cases.

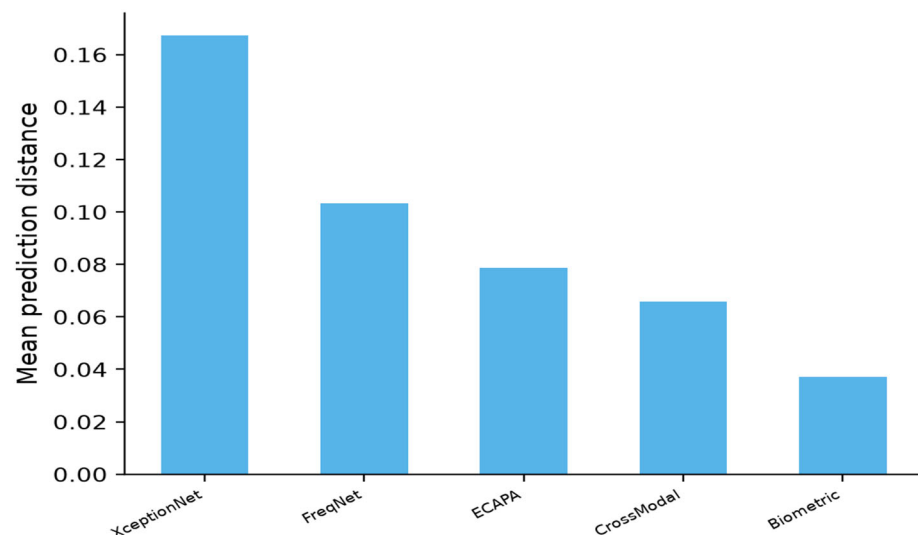


Figure 5. Mean prediction distance by agent; lower values indicate more confident, accurate predictions.

Third, score correlation is dominated by the shared label signal: the Cauchy–Schwarz bound forces at least 0.65 for the two strongest agents (Figure 6). Ensemble gain depends on error correlation: at the agents’ own threshold of 0.5 the largest is 0.307 (cross-modal–biometric), then 0.287 (visual–biometric) and 0.255 (the audio pair); several pairs are slightly negative. Cross-modal and biometric-quality correlate at 0.890 in score yet share only eight errors (of 43 and 15).

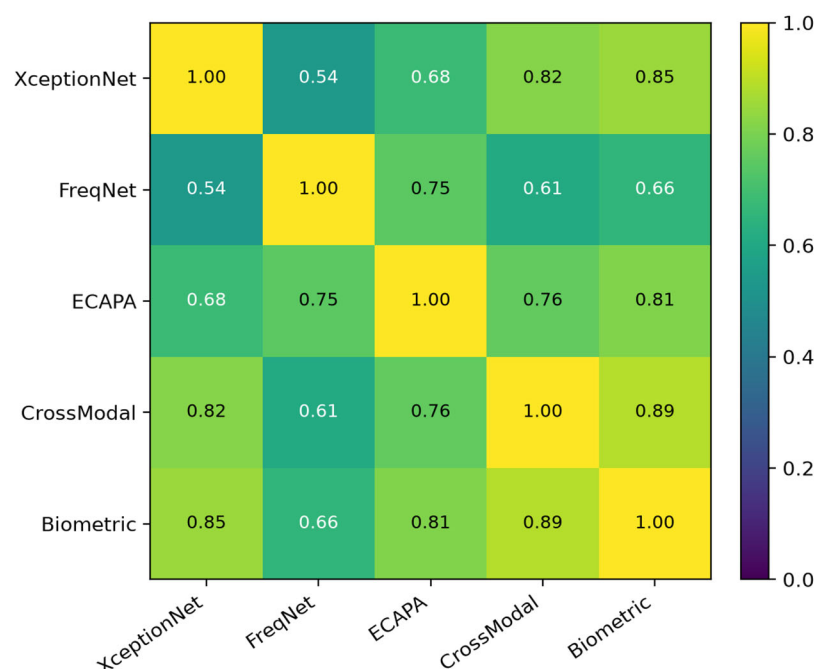


Figure 6. Agent score correlation heatmap on the PolyGlottFake test set (range 0.539–0.890); the error correlations that govern ensemble benefit are far lower (Section 4.3).

DeepfakeStack [37] varies architecture alone (Section 2.3); diversity here spans modality and forensic role: every audio–visual error correlation is at or below 0.26. This is a key design finding: architectural and feature diversity within the ensemble matters for robust performance, even when agents target overlapping modalities.

4.4. Ablation Studies

Systematic ablations quantify each component's contribution (Table 9). With equal weights no single agent dominates: removing FreqNet, the cross-modal agent or Biometric-Quality raises the error count from two to three, every added error a false alarm, while removing the visual agent or ECAPA-TDNN leaves it at two; no removal is significant on 2162 samples (Section 4.5.3), and the leave-one-out AUC-ROC stays at 1.000 except for ECAPA-TDNN (0.99997) and Biometric-Quality (0.99998). The visual agent is retained for the spatial evidence it contributes: its saliency maps are what escalation delivers to the analyst, and its removal costs nothing in-distribution only because the other agents cover the same clips, the redundancy the design intends; under the forgery families of MAVOS-DD it is the agent that transfers least (Section 4.11).

The gap between two-agent single-modality configurations ("Audio only" 98.84%; "Visual only" 99.58%) and the full ensemble (99.91%) quantifies the multi-modal fusion benefit at 0.3 to 1.1 percentage points. The single-best-agent configuration (Biometric-Quality alone at 99.63%, 8 errors) confirms that even the strongest individual agent falls short of the ensemble.

Table 9. Ablation study: agent-removal impact on the full test set, evaluated at $\tau = 0.35$. Weights of remaining agents are renormalised to sum to 1 in each configuration.

Configuration	Agents	Acc. (%)	Misc.	Δ (pp)
Full 5-agent ensemble	All five	99.91	2	—
Remove XceptionNet (Visual)	4 agents	99.91	2	0.000
Remove FreqNet (Audio)	4 agents	99.86	3	−0.046
Remove ECAPA-TDNN	4 agents	99.91	2	0.000
Remove Cross-Modal (Lip-Sync)	4 agents	99.86	3	−0.046
Remove Biometric-Quality	4 agents	99.86	3	−0.046
Top-3 (Phase-1 only)	Bio., ECAPA, CM	99.91	2	0.000
Audio only	FreqNet + ECAPA	98.84	25	−1.064
Visual only	Xcept. + Bio.	99.58	9	−0.324
Single best agent	Biometric-Quality	99.63	8	−0.278

4.5. Hyperparameter Sensitivity Analysis

Sensitivity was explored along three axes: the decision threshold that turns the aggregate score into a real/fake classification, the disagreement threshold that governs the rate of escalation to all five agents, and the statistical uncertainty of the resulting estimates.

4.5.1. Decision Threshold (τ) Sensitivity

Table 10 reports decision-threshold sensitivity on the validation partition over $\tau \in [0.20, 0.60]$; the test set is read once at $\tau = 0.35$. Under equal weights the partition is never classified perfectly: one fake scores 0.386 and one real 0.414. Between 0.277 and 0.386 the single error is that real, flagged; above 0.414 and up to 0.595 it is that fake, missed; in the gap between, both; accuracy stays above 98.3% across the window. The operating point of 0.35 is taken from the lower band so that no validation fake is missed, the choice that minimises the false-negative rate, and sits 0.074 above the second-highest real and 0.036 below the lowest fake; the conventional 0.5 would instead miss one validation fake. This score-level separability is why threshold-independent AUC-ROC and AP are the primary measures. Errors are asymmetric: below the gap they are false positives, above it false negatives, a direct consequence of near-separable score distributions.

Table 10. Decision-threshold robustness check on the validation partition (119 real, 119 fake), obtained by sweeping τ over the saved aggregate scores; the test set is read out once, at the operating point $\tau = 0.35$.

τ	Acc. (%)	FPR (%)	FNR (%)	Misc.
0.20	98.32	3.36	0.00	4
0.25	99.16	1.68	0.00	2
0.30	99.58	0.84	0.00	1
0.35	99.58	0.84	0.00	1
0.40	99.16	0.84	0.84	2
0.45	99.58	0.00	0.84	1
0.50	99.58	0.00	0.84	1
0.55	99.58	0.00	0.84	1
0.60	99.16	0.00	1.68	2

4.5.2. Disagreement Threshold Sensitivity

Table 11 reports the Phase-2 trade-off: accuracy is invariant (99.91%, two errors) for $\tau_{\text{disagree}} \in [0.20, 0.50]$, since any Phase-1 verdict split forces escalation; the default $\tau_{\text{disagree}} = 0.30$ gives full-ensemble accuracy at 3.09 agents per sample. The disagreement threshold is therefore a compute-cost control rather than an accuracy parameter, and 0.30 remains the released default.

Table 11. Disagreement-threshold impact on adaptive deployment on PolyGlottFake at $\tau = 0.35$. Escalation rate depends on τ_{disagree} and the Phase-1 score distribution, not on the decision threshold.

τ_{disagree}	Escalation (%)	Acc. (%)	Misc.	Avg. Agents
0.20	4.72	99.91	2	3.09
0.25	4.39	99.91	2	3.09
0.30	4.39	99.91	2	3.09
0.35	4.39	99.91	2	3.09
0.40	4.39	99.91	2	3.09
0.50	4.39	99.91	2	3.09

4.5.3. Statistical Uncertainty and Paired Significance Tests

All metrics are point estimates on one test partition ($n = 2162$); intervals and tests below use the stored per-agent scores (percentile bootstrap, 10,000 resamples, seed 42; paper_artifacts/task_21_statistical_treatment.py).

At $\tau = 0.35$ the two errors are both false positives: the Clopper–Pearson 95% interval on the 99.91% accuracy is [0.9967, 0.9999] and 95% bootstrap intervals are [0.9977, 1.000] (accuracy) and [1.000, 1.000] (AUC-ROC); the specificity is 116/118 on the 118 real samples and the sensitivity is 2044/2044 (no separate intervals). Clips from one source video are not independent, so the intervals were also computed over the 709 test source videos: [0.9976, 1.000] for accuracy and [1.000, 1.000] for AUC-ROC, against [0.9977, 1.000] and [1.000, 1.000] clip-level; the two errors come from different sources, so clustering does not widen them. On the source-disjoint partition the source-clustered (84 sources) accuracy intervals are [0.960, 0.990], [0.971, 0.995] and [0.982, 1.000] for seeds 42, 43 and 44, within 0.1 points of the clip-level intervals.

Leave-one-agent-out ablations re-aggregate the same stored scores, so exact McNemar tests apply: no single removal is significant. FreqNet (2 against 1, $p = 1.000$), Cross-Modal (1 against 0, $p = 1.000$), Biometric-Quality (1 against 0, $p = 1.000$), XceptionNet and ECAPA-TDNN (0 against 0, $p = 1.000$), so the ensemble’s margin over any four-agent subset is at most one sample, as equal weighting would suggest.

4.6. Training and Validation Dynamics

Training histories for every agent, including the validation-selected fine-tunes of FreqNet, the cross-modal agent and ECAPA-TDNN (Section 3.3), are given in Table S6 and Figure S1; each retained checkpoint is its agent's best-validation epoch.

Table 12 adds two over-fitting checks: the development-to-test accuracy gap and recall by synthesis method; a wide spread marks generator fitting. Biometric-Quality (−0.1-point gap, 0.9-point recall spread), Cross-Modal (−1.4, 1.1) and the frame-averaged visual agent (−0.3, 1.8) show no held-out degradation. The fine-tuned FreqNet scores 92.9% on validation and 93.6% on test, with recall from 100.0% (MicroTTS) to 31.6% (Vall-E-X), a 68.4-point spread (the released checkpoint spanned 73.4 points); ECAPA-TDNN's −0.3-point gap hides a 29.1-point spread on the same subset.

Leave-one-out re-aggregation of the stored scores needs no retraining: dropping FreqNet adds one error and leaves the AUC-ROC at 1.000; removing ECAPA-TDNN adds none, Biometric-Quality adds one. The most vocoder-fitted agent is thus a retained contributor: its errors fall on methods the others classify correctly, and its frequency-domain evidence separates samples near the boundary (Section 4.3). Its contribution is verified only in distribution; under shift it should fail first (Sections 4.11 and 5).

Table 12. Per-agent generalisation on the PolyGlottFake test set at the agents' own threshold of 0.5; development figures are the validation accuracies of the released agents, so the gaps are indicative. Recall (%) is given per synthesis method (651 MicroTTS, 338 Tacotron, 470 Bark, 506 XTTS and 79 Vall-E-X clips); the last row is the fused system at $\tau = 0.35$.

Agent	Dev. acc.	Test acc.	Gap (pp)	MicroTTS	Tacotron	Bark	XTTS	Vall-E-X
Visual (XceptionNet)	96.2	96.5	−0.3	95.2	97.0	97.0	96.8	96.2
Audio (FreqNet)	92.9	93.6	−0.7	100.0	100.0	99.8	84.0	31.6
Audio Forensics (ECAPA-TDNN)	97.5	97.8	−0.3	100.0	100.0	99.4	96.0	70.9
Cross-Modal (Lip-Sync)	96.6	98.0	−1.4	98.2	98.5	98.1	97.8	97.5
Biometric-Quality	99.2	99.3	−0.1	99.1	99.1	99.6	99.6	98.7
System ($\tau = 0.35$)	99.6	99.9	−0.3	100.0	100.0	100.0	100.0	100.0

4.7. Explainability Analysis

The explainability artefacts produced by the framework provide complementary forensic evidence across modalities. Figure 7 shows a composite output for all five agents on a single PolyGlottFake sample.

Visual GradCAM: XceptionNet saliency maps consistently highlighted the mouth and periorbital regions across correctly classified deepfakes, with secondary attention to forehead regions exhibiting excessive smoothness. The mouth concentration is consistent with PolyGlottFake's lip-sync manipulation methods (Wav2Lip, video retalking).

Audio GradCAM: FreqNet GradCAM overlays on Mel spectrograms highlighted compression artefacts from Wav2Lip and unnatural separation between frequency bands. A limitation of frequency-domain explanations is the difficulty of translating them back to specific temporal moments.

Cross-modal attention: The cross-attention heatmaps provide the most temporally informative explanations. For deepfakes, attention peaks appeared at frames with high audio-visual desynchronisation, typically at plosive consonants and fricatives. For real videos, attention was more uniformly distributed. The green border thickness on output frames was linearly correlated with attention values.

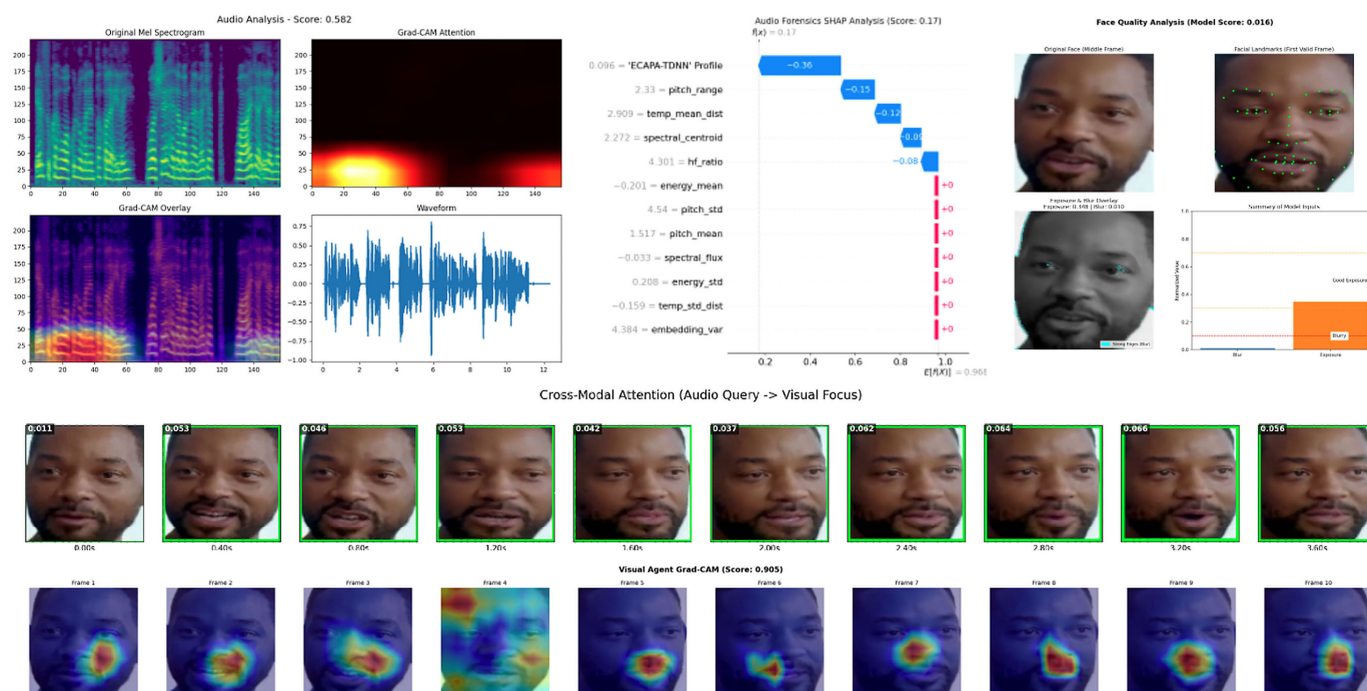


Figure 7. Composite explainability outputs for all five agents on a single PolyGlottFake sample (clockwise from top left): Audio FreqNet GradCAM on Mel spectrogram; ECAPA-TDNN SHAP waterfall plot; biometric-quality facial landmarks and quality metrics; cross-modal attention with green border intensity correlated with audio-visual desynchronisation; and Visual XceptionNet GradCAM highlighting the mouth region.

SHAP waterfall plots: The 192-dimensional speaker embedding features were the dominant predictors, followed by spectral centroid and pitch statistics, capturing unnatural vocal variations characteristic of voice cloning.

Fidelity and stability. Table 13 scores the three map-producing explanations on 60 test clips (30 authentic, 30 manipulated). Both GradCAM maps are faithful where the system detects a forgery: on clips scored as manipulated, deletion beats a random ordering on 26 of 32 clips for the visual map (0.357 against 0.543; exact sign test, $p = 0.0005$) and on 32 of 32 for the spectral map (0.058 against 0.607, $p < 0.001$). Against a location-matched control, with the same map rolled to a random position, the visual map is faithful on both classes, winning on 41 of 60 clips (0.530 against 0.597, $p = 0.006$) and on 21 of 28 scored as authentic ($p = 0.013$); a pixel-random ordering is a weaker test on authentic clips (9 and 12 of 28 for the two maps) because its speckle mask adds about ninety times more blur-to-sharp seam, which the detector reads as manipulation. The cross-modal attention is faithful in rank (score drop of 0.115 against 0.018 random and 0.004 least attended), though its top frame survives a flip in only 57% of clips, and the visual map is stable under JPEG ($q = 0.59$) but not under noise ($q = 0.33$), so both are read at the level of regions rather than pixels. SHAP attributions are not scored, and no metric uses manipulation masks, which the dataset lacks.

Table 13. Explanation fidelity and stability on 60 PolyGlottFake test clips (means; released system). Deletion and insertion follow the RISE protocol against a random ordering of the same input; for the cross-modal attention the three most-attended frames are replaced by the clip’s mean frame; stability is the Spearman correlation of the map recomputed under the named perturbation; the location-matched control is the same map rolled to a random position.

Explanation	Test	Map	Random Order	Stability (Spearman ρ)
Visual GradCAM (XceptionNet)	Deletion AUC (lower is better)	0.530	0.595	0.33 noise; 0.59 JPEG 60
	Insertion AUC (higher is better)	0.671	0.594	
	Deletion AUC, location-matched control (lower is better)	0.530	0.597	
Spectral GradCAM (FreqNet)	Deletion AUC (lower is better)	0.465	0.736	0.47 30 dB noise; 1.00 gain 0.7
	Insertion AUC (higher is better)	0.851	0.733	
Cross-modal attention Score drop, three frames removed		0.115 (most attended)	0.018 random; 0.004 least attended	0.64 flip; 0.37 noise

Natural-language reports: the Figure 8 report identifies the speaker, flags the English-to-Russian dub, and returns DEEPFAKE from unanimous first-wave scores (0.991, 0.982, 0.989; spread 0.004, Phase-1 aggregate 0.987).

Only the contextual plausibility analysis is not detector-derived; it rests on the report model’s unverifiable world knowledge and is a reporting-stage limitation (Section 5); the verdict is fixed against $\tau = 0.35$ before reporting.

Example deepfake analysis report

Digital forensics report: en_37_to_ru_Xfts_label_fake.npz

Executive summary

The system’s final verdict is DEEPFAKE with an aggregate score of 0.987. This verdict is justified by the aggregate agent score exceeding the deepfake decision threshold ($\tau = 0.35$), indicating a high likelihood of tampering.

Model analysis

The contributing agent scores are:

First-wave agents

Face Quality: 0.9911 — High, indicating artifacts

Audio Forensics (ECAPA): 0.9819 — High, indicating artifacts

Cross-Modal (Lip-Sync): 0.9891 — High, indicating artifacts

These scores collectively suggest a deepfake, with all three first-wave agents detecting significant artifacts.

Contextual plausibility analysis

- **Identity vs. content mismatch:** The identified person, Will Smith, seems out of character speaking in Russian.
- **Language mismatch:** Will Smith is not known to speak Russian fluently, making the language mismatch suspicious.
- **Overall scenario plausibility:** Implausible due to mismatched language and content.

Final verification

The system’s final verdict is DEEPFAKE, which matches the provided Ground Truth Label of Fake. Therefore, the system is CORRECT in its detection.

Deployment phases

Phase 1, first wave: Face Quality, Audio Forensics (ECAPA), Cross-Modal (Lip-Sync)

Phase 2: not deployed

Reason: First-wave verdicts unanimous, disagreement level below threshold ($0.004 \leq 0.3$).

Figure 8. Example forensic report generated by the multi-agent framework for a PolyGlottFake sample.

4.8. Qualitative Analysis: Best, Average, and Worst-Case Outputs

Best case (PolyGlottFake, $\approx 96\%$ of samples). All Phase 1 agents converge on high scores and unanimous verdicts: on nine in ten of these fakes ECAPA-TDNN scores at least 0.90, Cross-Modal at least 0.89 and Biometric-Quality at least 0.93. GradCAM maps consistently highlight the mouth and periorbital regions. The forensic report correctly identifies speakers, detects language mismatches, and flags identity–content inconsistencies. These samples are resolved entirely by Phase 1.

Average case (borderline, $\approx 4\%$ of samples). Phase 2 is triggered because the first-wave verdicts split or their scores spread. The most common escalation pattern involves ECAPA-TDNN scoring low (authentic) while the Cross-Modal and Biometric-Quality agents score high (deepfake), which is typical of samples with convincing audio but visible face-track artefacts. Phase 2 agents resolve these with additional spatial and frequency-domain evidence.

Worst case (YouTube misses). Thirteen of the 16 AI-avatar clips are missed at $\tau = 0.35$: seven lie within 0.10 of the boundary, the highest missed aggregate is 0.341 and four score below 0.20. The three detections are the HeyGen-viewer clip (0.539), on which the biometric-quality and visual agents fire (0.96 and 0.83), and two digital-twin clips at 0.385 and 0.351 carried by the audio agents; on the misses ECAPA-TDNN fires on nine of the 13 while the biometric-quality and cross-modal agents stay silent (Figure 9), and under equal weights the audio agents alone rarely carry the aggregate past the threshold; diffusion-based avatars likely exceed the fidelity of PolyGlottFake’s translation-based manipulations.



Figure 9. Worst-case explainability outputs from the YouTube evaluation (a missed AI-avatar clip, aggregate 0.298 against $\tau = 0.35$). The visual GradCAM wanders between frames and the cross-modal attention collapses onto the first frame, so neither isolates a consistent manipulation region, which is the pattern of a distribution mismatch between training and test data; the audio explanations lean towards fake (ECAPA-TDNN SHAP profile 0.83, FreqNet GradCAM 0.56) but the silent biometric-quality (0.00) and cross-modal (0.03) agents hold the aggregate below the threshold.

4.9. Adaptive Deployment Efficiency

Table 14 summarises deployment statistics. On PolyGlottFake the first wave handles approximately 96% of inputs; Phase 2 fires on only 4.4% of samples. The escalation rate itself serves as a distribution-shift indicator: on the YouTube benchmark, this figure rises to 32.0% and on MAVOS-DD to 47.1%, seven-fold and eleven-fold increases that signal domain mismatch and could trigger retraining or human-in-the-loop review in production deployment.

Table 14. Adaptive deployment statistics across datasets.

Dataset	N	Ph. 1 Only	Ph. 1 + 2	Esc. (%)	Acc. (%)
PolyGlottFake (full)	2162	2067	95	4.39	99.91
PolyGlottFake source-disjoint	387	347	40	10.34	98.45
YouTube custom	50	34	16	32.00	74.00
MAVOS-DD	2861	1514	1347	47.08	36.35

4.10. Baseline Comparison on the English Subset

Table 15 reports the English-subset results that informed the choice of baselines, single-run point estimates from our implementations: the hybrid CNN-Transformer reached 99.1%, and full-face input 99% against 98% for mouth crops.

Table 15. Model comparison on the PolyGlottFake English subset; single-run point estimates from our implementations. Epochs are per-model training budgets (Section 3.7); splits per Section 3.1.

Model	Epochs	Params (M)	Test acc.
GenConViT VAE	5	140	95%
GenConViT AE (ED)	10	130	98%
LIPINC-V2	10	4.9	90.2%
Custom ViT (late fusion)	15	181	87.6%
CNN-Transformer (hybrid)	20	6.42	99.1%

4.11. External and Real-World Generalisation

To test transfer beyond PolyGlottFake, the released system was scored unchanged (equal weights, $\tau = 0.35$) on MAVOS-DD [53], a public audio-visual benchmark of YouTube-sourced speakers in eight languages whose forgeries differ in kind from the training data: PolyGlottFake dubs synthesised speech onto authentic footage and re-renders the mouth, whereas MAVOS-DD forges the video by whole-face animation (EchoMimic, Memo, Sonic), reenactment (LivePortrait) or face swap (Roop, HifiFace, Inswapper), mostly with authentic speech, and converts the voice (kNN-VC, FreeVC, OpenVoice, XTTS v2) in part of the clips. A seeded stratified sample of the official test split (the released manifest lists each clip) gave 2861 scored clips, 2550 forged and 311 authentic from 2655 source videos, of which 1540 forge only the video, 695 both streams and 315 only the voice; every clip was also re-encoded at constant rate factor (CRF) 23 and CRF 40, and intervals are 95% bootstraps over source videos.

Table 16 shows where the boundary lies. Wherever the speech is forged (1010 clips) the system reaches an AUC-ROC of 0.788 (0.754–0.819) and its audio agents alone 0.904 (0.884–0.920), including on conversion methods absent from training (ECAPA-TDNN 0.863 on kNN-VC and 0.926 on OpenVoice); where only the video is forged (1540 clips) it does not transfer, 0.444 (0.408–0.481), because the face-based agents learned the cues of the partial-forgery family: the cross-modal agent looks for the desynchrony that dubbing leaves and Xception and Biometric-Quality for the seams of mouth re-rendering, neither of which the animators produce, while both remain informative on Roop and LivePortrait

(per class, Table S7). Over all 2550 forged clips the AUC-ROC is 0.580 (0.545–0.612) and the balanced accuracy is 56.5% (54.0–58.9%), above chance, but at $\tau = 0.35$ recall is 30.7% at a specificity of 82.3%, so raw accuracy on this 8.2:1 sample is 36.4%; by language, only Mandarin, Russian, Spanish and Romanian are above chance.

The first wave disagrees on 47.1% of clips against 4.4% in distribution, so the escalation rate flags the shift without labels (Table 14), and here escalation pays: the five-agent system makes 166 fewer errors than the Phase-1 trio (1821 against 1987; exact McNemar $p < 0.001$). Compression does not degrade the ranking (AUC-ROC 0.579, 0.597 and 0.674 for the original clips and at CRF 23 and 40, with the audio agents unchanged) but breaks the operating point at CRF 40, with recall rising from 30.6% to 75.6% while specificity falls from 82.3% to 38.3%. Most of the loss is recoverable within the framework: re-fitting only the fusion weights on MAVOS-DD labels (a diagnostic, not a transfer result) reaches 0.710 (0.680–0.739), though no single weight vector suits every family (leaving out the class under test degrades Roop, 0.752 to 0.447), and one EfficientNet-B0 agent trained in 30 min on MAVOS-DD's training split (four generators; 1054 training and 274 validation clips; no source video shared with the test sample), added to the unchanged five at equal weights and unchanged τ , raises the AUC-ROC from 0.580 to 0.740 (0.703–0.774) with no class degraded (Table 16 and, per class, Table S7), including from 0.568 to 0.735 on the three generators it never saw. Alone it reaches 0.857, above the fusion, so under shift equal weighting is the binding constraint (Section 5); this is a within-benchmark result, showing that a missing family can be covered without retraining the existing agents, not that the added agent generalises beyond MAVOS-DD.

Table 16. MAVOS-DD by forged modality and compression tier (released system, equal weights, $\tau = 0.35$; each group against the 311 authentic clips): system AUC-ROC with 95% source-clustered bootstrap intervals, the AUC-ROC of the six-agent fusion with the added agent of Section 4.11 (not scored under compression), recall at $\tau = 0.35$, and per-agent AUC-ROC. Results per forgery class are in Table S7.

Subset	Forged Clips	System AUC-ROC (95% CI)	With Added Agent	Recall	Xception	FreqNet	ECAPA-TDNN	Cross-Modal	Biometric-Quality
<i>By forged modality</i>									
Video only	1540	0.444 (0.408–0.481)	0.662	12.3%	0.386	0.493	0.582	0.371	0.405
Video and speech	695	0.828 (0.795–0.860)	0.921	69.1%	0.310	0.873	0.954	0.384	0.329
Speech only	315	0.697 (0.651–0.739)	0.720	36.5%	0.507	0.634	0.861	0.552	0.475
All forged clips	2550	0.580 (0.545–0.612)	0.740	30.7%	0.380	0.614	0.718	0.397	0.393
<i>By compression tier (2856 clips scored in all three)</i>									
Original	2545	0.579 (0.546–0.610)	—	30.6%	0.379	0.614	0.717	0.396	0.392
CRF 23	2545	0.597 (0.564–0.627)	—	34.9%	0.397	0.617	0.713	0.449	0.415
CRF 40	2545	0.674 (0.645–0.701)	—	75.6%	0.467	0.617	0.713	0.633	0.495

A second, smaller evaluation set comprises 50 clips from eleven public YouTube videos (BBC News and AI-avatar content), clipped by the released extraction script's gates: MTCNN confidence ≥ 0.95 , a 100-source-pixel face minimum, a sharpness floor, and ten-frame identity continuity. The released manifest records the replacement. The accuracy is 74.00% (95% Clopper–Pearson interval 59.7–85.4%); Table 17 gives the $\tau = 0.35$ confusion matrix.

There are no false positives (no real clip scores above 0.22); three of the 16 avatar clips are detected, the HeyGen-viewer clip at 0.539 and two digital-twin clips at 0.385 and 0.351, and thirteen are missed, seven of them within 0.10 of the boundary; Figure 10 shows the two nearest misses, at 0.341 and 0.334. The ranking survives the shift far better than the operating point (AUC-ROC 0.98): on the digital-twin clips the audio agents order the

avatars correctly but, at 0.40 of the weight, seldom lift them past the threshold; escalation rises from 4.4% to 32.0%.

Table 17. Confusion matrix on the custom YouTube evaluation set (n = 50 clips: 34 real, 16 fake; evaluated at $\tau = 0.35$).

	Pred. Real	Pred. Fake
Actual Real (34)	34	0
Actual Fake (16)	13	3



Figure 10. Frames from the two nearest misses among the AI-avatar clips, aggregate scores 0.341 and 0.334 against $\tau = 0.35$. Diffusion-generated content, absent from PolyGlottFake, renders too cleanly for the biometric-quality and cross-modal agents to fire.

4.12. Inference Cost

Table 18 reports batch-1 latency and accelerator memory for all agents and three transformer baselines: five warm-up passes, 30 timed samples of 10 runs, device-synchronised, separate processes, preprocessing excluded. Baselines are rebuilt to published specification with random weights; the Hybrid CNN-Transformer rebuild is slightly larger (7.49 M against 6.42 M parameters).

Table 18. Inference cost on Apple Metal (MPS; Apple, Cupertino, CA, USA), batch size 1, float32. Parameter memory is hardware-independent; peak memory on unified memory is not comparable with discrete-GPU figures; parallel latency assumes concurrent dispatch; visual and biometric rows are single-crop forwards; † rebuild smaller than published.

Model/Configuration	Params (M)	Param mem. (MB)	Peak mem. (MB)	Latency (ms)
Visual (XceptionNet)	23.96	91.4	129.6	11.9
Audio (FreqNet)	1.85	7.1	117.9	25.1
Cross-Modal (Lip-Sync)	6.13	23.4	249.6	23.0
Biometric-Quality	4.36	16.6	41.9	6.8
Audio Forensics (ECAPA-TDNN)	20.81	79.4	—	35.7
Phase 1 only (three agents, 95.6% of samples)	31.30	119.4	—	35.7 par./65.5 seq.
Phase 1 + 2 (five agents, 4.4% of samples)	57.11	217.8	—	60.9 par./102.6 seq.
Expected under adaptive deployment	—	123.7	—	36.8 par./67.1 seq.
Hybrid CNN-Transformer	7.49	28.6	41.4	8.6
GenConViT AE †	113.57	433.2	463.7	28.3
Late-fusion ViT †	128.43	489.9	497.5	32.2

With Phase 2 firing on 4.4% of samples, expected parameter memory is 123.7 MB, or 119.4 MB for Phase 1 alone, against the large baselines' 433.2 MB and 489.9 MB (both lower bounds). Concurrent dispatch makes peak allocation additive (at least 291.6 MB for Phase 1, 539.1 MB for all five) but cuts expected latency from 67.1 ms sequential (as released) to 36.8 ms, needed for latency parity.

The hybrid CNN-transformer baseline is cheaper on both axes (28.6 MB, 8.6 ms); its 98.1% is a single-run point estimate, so no accuracy difference is established, and what the framework adds is per-modality explanations and an escalation audit trail, at roughly four times the parameter memory and four to eight times the latency.

5. Discussion

Five findings warrant emphasis. First, the five-agent ensemble reaches 99.91% accuracy and threshold-free AUC-ROC of 1.000 on PolyGlottFake with equal weights at $\tau = 0.35$, and 98.4% on the source-disjoint partition (97.7% without test sources that may share an inferred identity with training), where every error is a false alarm on an authentic clip; the baselines' figures are single-run point estimates from our implementations and support no claim of superiority (see Threats to Validity). The gap between the single-modality configurations (98.8 to 99.6%) and the full ensemble, and between the single-best agent (Biometric-Quality alone, 99.63%) and the ensemble, is consistent with a gain from multi-modal fusion, but none of these in-distribution differences is statistically significant on 2162 clips (Section 4.5.3); the ensemble's contribution is established under distribution shift, where the five agents make significantly fewer errors than the first-wave trio (Section 4.11).

Second, the inter-agent correlation analysis (Section 4.3) clarifies what architectural variation in a multi-agent system actually looks like. The five agents include two audio agents (FreqNet and ECAPA-TDNN) and two visual agents (XceptionNet and Biometric-Quality). Their errors are far less correlated than their scores: the four largest error correlations lie between 0.24 and 0.31 and every other pair falls below 0.03. ECAPA-TDNN's embedding-based approach captures different vocal features from FreqNet's frequency-domain analysis, and XceptionNet's deeper spatial features complement the Biometric-Quality agent's quality-channel input. This suggests that architectural and feature diversity within the ensemble matters more than agent count.

Third, the integrated explainability differs from current practice. Existing systems use GradCAM, SHAP, or LIME as post hoc analytical tools applied to a finished model. This framework generates modality-specific visual explanations alongside a synthesised natural-language report, providing both machine-grounded evidence and a human-readable rationale. The systematic XAI comparison (Section 3.6) found that applying multiple methods together produced noticeable discrepancies that could confuse downstream readers. This motivated the selection of the most computationally efficient and interpretable method for each modality.

Fourth, the hyperparameter sensitivity analysis (Section 4.5) shows that the system is robust across the swept window: the validation accuracy stays above 98.3% throughout the 0.20–0.60 window, the partition makes a single error for every τ from 0.277 to 0.386 (a false alarm) and above 0.414 up to 0.595 (a missed fake), and $\tau = 0.35$ is taken from the lower band to minimise missed fakes. The disagreement threshold changes cost, not accuracy (Table 11): Phase 2 does not significantly improve in-distribution accuracy, and its supported contributions are the additional forensic evidence it renders, an escalation rate that rises under domain shift and significantly fewer errors on MAVOS-DD (Sections 4.1, 4.9 and 4.11).

Fifth, the external evaluation (Section 4.11) shows what the modular design offers when the forgery family changes: where a monolithic detector falling to an AUC-ROC of 0.580 gives no indication of why, the per-agent scores show that the audio agents transfer, attribute the loss to the face-based agents and locate the gap in video forged with authentic speech, and one added agent for that family raised the AUC-ROC to 0.740 without retraining the existing five. The remaining gap is fusion rather than coverage: the added agent alone scores 0.857, above the 0.740 of the six-agent equal-weight fusion, so added agents

would need family-aware routing, which the present fixed-weight orchestrator does not provide (Section 6).

Threats to Validity

Internal validity. Near-perfect accuracy on PolyGlottFake may partly reflect dataset-specific artefacts; the dataset's manipulation methods are known and finite, and the ensemble may have learned dataset-specific patterns. The test partition is also heavily imbalanced, 118 real clips against 2044 fake, so a rule that labelled every clip fake would already score 94.54%; accuracy is a weak summary at this prevalence, and the threshold-free AUC-ROC and average precision of Section 4.1, together with the per-class counts, carry the evidence. Both errors at $\tau = 0.35$ fall on the real class, which holds only 118 clips, so the specificity is 98.31% with a 95% interval of 94.0% to 99.8%: the false-alarm behaviour of the system is estimated far less precisely than the headline accuracy suggests. The operating threshold is itself fixed by single clips on a 238-clip validation partition, one real scoring 0.414 and one fake 0.386 (Section 4.5.1), and 0.30 satisfies the same criterion, so the choice rests on a thin margin in a small sample rather than on a stable optimum. Two agents are fitted narrowly to the synthesis methods they were trained on: FreqNet's recall spans 68.4 percentage points across methods and ECAPA-TDNN's 29.1 (Table 12), so both should degrade first on an unseen vocoder, although the cross-modal and biometric agents cover those clips here. Seed 42 reproduces checkpoint inference but not GPU training. On the source-disjoint partition, retrained under three seeds, the accuracy is 98.4% (standard deviation of 0.8 points) and every error is a false alarm on one of the 83 authentic clips (Section 4.1): the clip-level 99.91% measures within-source generalisation and the source-disjoint figure generalisation to held-out facial-identity clusters, which are inferred rather than labelled, so some overlap may remain; without the 48 test sources flagged as possible overlaps it is 97.7% (Section 3.1), with a wide specificity interval on the small real side. The baseline figures of Tables 7 and 15 are point estimates from our implementations: the baselines were trained once, outside the released code, only epoch budgets were recorded and per-sample predictions were not retained (Section 3.7), so they may understate tuned baseline performance, no paired test is possible, and they are not repeated on the source-disjoint partition. Further models could be trained on different datasets, particularly diffusion-specific artefacts, to improve performance on in-the-wild benchmarks.

External validity. The YouTube evaluation is an internal distribution-shift measure, not in-the-wild evidence: 50 author-constructed clips from eleven videos, withheld for copyright but regenerable from the released source identifiers and extraction script. Its fake side is narrower than the clip count implies: the 16 avatar clips come from four source videos and show renderings of a small number of presenters, so the set probes one family of generator behaviour rather than the diversity of synthetic media in circulation. Recall on that side is three of 16 (18.8%, 95% interval of 4.0% to 45.6%), an interval far too wide to rank generators or to support a general claim about avatar content, and the reported accuracy is carried mainly by the 34 real clips, none of which is misclassified. On Deepfake-Eval-2024 [55] off-the-shelf open-source detectors score AUC of 0.43 to 0.63, and GenConViT, the strongest published baseline of Table 7, falls from 0.96 on its own test data to 0.63. MAVOS-DD (Section 4.11) sharpens the pattern: the unchanged system detects forged speech (AUC-ROC of 0.788) but not video forgeries with authentic speech (0.444), so its overall 0.580 is above chance yet far below in-domain performance.

Construct validity. Section 4.7 scores the fidelity and stability of the three map-producing explanations, and finds both GradCAM maps faithful where the system detects a forgery, the visual map also against a location-matched control on both classes, with stability under noise being the remaining limit; localisation is not scored, because no ground-truth manipulation masks accompany these clips, and the perceived interpretability of the

reports has not been validated with domain experts (forensic analysts, content moderators, end users). The reports' contextual inferences are unevaluated, though they cannot affect classification.

Calibration. The aggregate score is a weighted combination of five independently trained sigmoid outputs rather than a single BCE-trained classifier. It is discriminative but not probabilistically calibrated: the expected calibration error with 10 equal-width bins is 0.090. A score of, say, 0.6 should not be read as “60% probability of being a deepfake”; it should be read only relative to the decision boundary. Downstream applications requiring calibrated probabilities would benefit from an explicit calibration step (Platt scaling or isotonic regression on a held-out split). No probabilistic claims about individual score values are made in this work.

Computational considerations. Training took around 520 GPU hours, so the framework remains more computationally intensive at training time than single-model alternatives, and the orchestration overhead must be amortised across inferences in production. Adaptive deployment mitigates inference cost: the Phase 1 agents total 31.3 M parameters (119 MB) and resolve around 96% of samples (Table 18), although sequential inference latency is roughly twice a monolithic pass and a smaller hybrid baseline is cheaper on both axes. Latencies were measured on Apple Metal rather than on a CUDA deployment GPU (Table 18), so the absolute timings are indicative and only the ratios between configurations should be read across hardware.

6. Conclusions

This paper has presented a multi-agent framework for explainable multi-modal deepfake detection. On the PolyGlotFake test set, a within-source measurement, it reaches 99.91% accuracy (two false alarms and no missed fake in 2162 clips) with a threshold-free AUC-ROC of 1.000, and on a partition disjoint in source video and inferred facial identity, it reaches 98.4% over three seeds (97.7% without test sources that may share an identity with training), with equal agent weights, $\tau = 0.35$ set on the validation partition, and a parameter budget (57.1 M, 13.3 M trainable) under half that of a single ViT baseline. The findings supported by the present experiments are fivefold: (i) a five-agent design whose complementary coverage is measured rather than assumed (every pairwise error correlation at or below 0.31); (ii) integrated explainability combining GradCAM, SHAP, attention mapping and natural-language reporting, with fidelity and stability measured (Section 4.7); (iii) adaptive orchestration that resolves around 96% of samples with three agents, where escalation does not significantly improve in-distribution accuracy but adds forensic evidence, rises seven- to eleven-fold under shift and significantly reduces errors on MAVOS-DD; (iv) ablation, sensitivity and measured inference-cost analyses; and (v) external evaluations on MAVOS-DD and a YouTube set, in which per-agent diagnostics attribute the loss under shift to video forgery families absent from training and one added agent raises the benchmark AUC-ROC from 0.580 to 0.740.

Future work follows from the stated limits: a learned, family-aware orchestration policy, potentially using reinforcement learning; expansion to include additional agents whose evidence decorrelates from the present ones, such as dense optical flow or the facial geometry and illumination cues of [21–23]; localisation against manipulation masks and user studies with forensic analysts; and evaluation on larger in-the-wild benchmarks. A tool-based extension, in which the detectors run in a sandbox behind a Model Context Protocol (MCP) server and a reasoning agent calls them according to the suspected forgery and reasons over their explanations, is outlined in Text S2.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/informatics13100163/s1>, Table S1: Clip accounting and class

composition of the source-disjoint partition by source language; Table S2: Manipulated clips of the source-disjoint partition by synthesis method and target language; Table S3: Checks of the inferred facial-identity clusters and the conservative test read-out; Table S4: Visual agent hyperparameter selection; Table S5: Audio baseline architecture selection; Table S6: Training dynamics of the individual agents; Figure S1: Per-epoch training and validation loss; Table S7: MAVOS-DD results by forgery class; Text S1: Agent implementation details; Text S2: A tool-based extension.

Author Contributions: Conceptualization, S.B.; methodology, S.B.; software, S.B.; formal analysis, S.B.; investigation, S.B.; data curation, S.B.; writing, original draft preparation, S.B.; writing, review and editing, S.B., P.S. and S.V.; visualization, S.B.; supervision, P.S. and S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The article processing charge (APC) was supported by the Imperial College London Open Access Fund.

Institutional Review Board Statement: This work used the publicly released PolyGlottFake academic dataset under its terms of release for academic research. The complementary YouTube evaluation used publicly available content for non-commercial academic research only; no video or clip is redistributed, face embeddings computed during extraction served only to enforce within-clip identity continuity and were not retained, and the figures reproduce only generated avatar faces.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in <https://github.com/saoirsebarry/multiagent-deepfake-detection> (accessed on 28 September 2026).

Acknowledgments: This work was carried out as part of the corresponding author's individual master's project on the Artificial Intelligence programme at Imperial College London, under the supervision of Susmitha Vekkot and Pancham Shukla. The authors acknowledge the use of generative AI assistants (Anthropic's Claude and Google's Gemini models) for code generation and debugging. In preparing this work and its revisions, the authors used Claude (Anthropic, San Francisco, CA, USA) under their direction for drafting and editing the manuscript, and implementing and running the released analysis, training and reproducibility scripts. All AI-assisted output was reviewed, verified against the released data and code, and edited by the authors, who take full responsibility for this publication. Generative AI was not used to fabricate, alter, or misrepresent research data or results, and AI tools are not listed as authors. The studied framework itself employs large language and vision-language models (Section 3); that usage is part of the studied system, not of this article's authorship.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AE	Autoencoder
AP	Average Precision
ASR	Automatic Speech Recognition
AUC	Area Under Curve
BCE	Binary Cross-Entropy
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CQT	Constant-Q Transform
ECAPA-TDNN	Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Network
EER	Equal Error Rate
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network

GradCAM	Gradient-weighted Class Activation Mapping
LFCC	Linear-Frequency Cepstral Coefficient
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LMM	Large Multi-Modal Model
MAS	Multi-Agent System
MFCC	Mel-Frequency Cepstral Coefficient
MTCNN	Multi-task Cascaded Convolutional Network
OOD	Out-Of-Distribution
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SOTA	State-Of-The-Art
SSM	State Space Model
VAE	Variational Autoencoder
ViT	Vision Transformer
VLM	Vision-Language Model
XAI	eXplainable Artificial Intelligence

References

1. Gandhi, K.; Kulkarni, P.; Shah, T.; Chaudhari, P.; Narvekar, M.; Ghag, K. A Multimodal Framework for Deepfake Detection. *J. Electr. Syst.* **2024**, *20*, 4772–4787. <https://doi.org/10.53555/jes.v20i10s.6126>.
2. Chesney, R.; Citron, D.K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. Law Rev.* **2019**, *107*, 1753.
3. Lao, Y.; Hirvonen, N.; Larsson, S. AI and authenticity: Young people's practices of information credibility assessment of AI-generated video content. *J. Inf. Sci.* **2026**, *52*, 975–989. <https://doi.org/10.1177/01655515251330605>.
4. Milmo, D. UK Engineering Firm Arup Falls Victim to £20m Deepfake Scam, 2024. Available online: <https://www.theguardian.com/technology/article/2024/may/17/uk-engineering-arup-deepfake-scam-hong-kong-ai-video> (accessed on 28 September 2026).
5. Deloitte. Deepfake Banking Fraud Risk on the Rise, 2024. Available online: <https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html> (accessed on 28 September 2026).
6. SecurityHero. State of Deepfakes Report, 2023. Available online: <https://www.securityhero.io/state-of-deepfakes/#key-findings> (accessed on 28 September 2026).
7. Amerini, I.; Barni, M.; Battiato, S.; Bestagini, P.; Boato, G.; Bruni, V.; Caldelli, R.; De Natale, F.; De Nicola, R.; Guarnera, L.; et al. Deepfake Media Forensics: Status and Future Challenges. *J. Imaging* **2025**, *11*, 73. <https://doi.org/10.3390/jimaging11030073>.
8. Zhou, Y.; Lim, S.N. Joint Audio-Visual Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021, pp. 14780–14789. <https://doi.org/10.1109/ICCV48922.2021.01453>.
9. Muppalla, S.; Jia, S.; Lyu, S. Integrating Audio-Visual Features For Multimodal Deepfake Detection. In *Proceedings of the 2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*; IEEE: New York, NY, USA, 2023, pp. 1–5. <https://doi.org/10.1109/URTC60662.2023.10534969>.
10. Deressa, D.W.; Mareen, H.; Lambert, P.; Atnafu, S.; Akhtar, Z.; Van Wallendael, G. GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer. *Appl. Sci.* **2025**, *15*, 6622. <https://doi.org/10.3390/app15126622>.
11. Khan, S.A.; Dang-Nguyen, D.T. Hybrid Transformer Network for Deepfake Detection. *arXiv* **2022**, arXiv:2208.05820.
12. Khormali, A.; Yuan, J.S. DFDt: An End-to-End DeepFake Detection Framework Using Vision Transformer. *Appl. Sci.* **2022**, *12*, 2953. <https://doi.org/10.3390/app12062953>.
13. Mansoor, N.; Iliev, A.I. Explainable AI for DeepFake Detection. *Appl. Sci.* **2025**, *15*, 725. <https://doi.org/10.3390/app15020725>.
14. Kaddar, B.; Fezza, S.A.; Akhtar, Z.; Hamidouche, W.; Hadid, A.; Serra-Sagristá, J. Deepfake Detection Using Spatiotemporal Transformer. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–21. <https://doi.org/10.1145/3643030>.
15. Silva, S.H.; Bethany, M.; Votto, A.M.; Scarff, I.H.; Beebe, N.; Najafirad, P. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Sci. Int. Synerg.* **2022**, *4*, 100217. <https://doi.org/10.1016/j.fsisyn.2022.100217>.
16. Govindu, A.; Kale, P.; Hullur, A.; Gurav, A.; Godse, P. Deepfake audio detection and justification with explainable artificial intelligence. *Int. J. Data Sci. Anal.* **2026**, *22*, 212. <https://doi.org/10.1007/s41060-026-01186-7>.

17. Anitha, A.; Saju, L.M.; Kamaraj, B. Deepfake Detection using XAI based Deep Fusion Models. In *Proceedings of the 2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*; IEEE: New York, NY, USA, 2025, pp. 526–531. <https://doi.org/10.1109/CICTN64563.2025.10932587>.
18. Rong, Y.; Leemann, T.; Nguyen, T.T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; Kasneci, E. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2104–2122. <https://doi.org/10.1109/TPAMI.2023.3331846>.
19. Raval, H.; Patel, M.; Parmar, S. A Review on Explainable AI for Deepfake Detection Leveraging Hybrid Deep Learning Techniques. *SPU-J. Sci. Technol. Manag. Res.* **2025**, *2*, 69–75. <https://doi.org/10.63766/spujstmr.24.000034>.
20. Wesselkamp, V.; Rieck, K.; Arp, D.; Quiring, E. Misleading Deep-Fake Detection with GAN Fingerprints. *arXiv* **2022**, arXiv:2205.12543.
21. Cheng, Z.; Chen, C.; Zhou, Y.; Hu, X. Mining Temporal Inconsistency with 3D Face Model for Deepfake Video Detection. In *Pattern Recognition and Computer Vision (PRCV 2023)*; Lecture Notes in Computer Science; Springer Nature: Singapore, 2023; Volume 14431, pp. 231–243. https://doi.org/10.1007/978-981-99-8540-1_19.
22. Tian, K.; Chen, C.; Zhou, Y.; Hu, X. Illumination Enlightened Spatial-temporal Inconsistency for Deepfake Video Detection. In *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2024, pp. 1–6. <https://doi.org/10.1109/ICME57554.2024.10687905>.
23. Zhang, G.; Liang, Y.; Tian, K.; Yi, J.; Alsolai, H.; Liu, M.; Hu, X. Leveraging Spatial-Temporal Illumination Features and Convolution-Transformer Hybrid Networks for Deepfake Video Detection. *IEEE Trans. Consum. Electron.* **2025**, *71*, 12479–12489. <https://doi.org/10.1109/TCE.2025.3624764>.
24. Patil, K.; Kale, S.; Dhokey, J.; Gulhane, A. Deepfake Detection using Biological Features: A Survey. *arXiv* **2023**, arXiv:2301.05819.
25. Seibold, C.; Wisotzky, E.L.; Beckmann, A.; Kossack, B.; Hilsman, A.; Eisert, P. High-quality deepfakes have a heart! *Front. Imaging* **2025**, *4*, 1504551. <https://doi.org/10.3389/fimag.2025.1504551>.
26. Zhang, B.; Cui, H.; Nguyen, V.; Whitty, M. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. *Sensors* **2025**, *25*, 1989. <https://doi.org/10.3390/s25071989>.
27. Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; Wei, Y. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning. *arXiv* **2024**, arXiv:2403.07240.
28. Yan, J.; Li, Z.; Wang, F.; He, Z.; Fu, Z. Dual Frequency Branch Framework with Reconstructed Sliding Windows Attention for AI-Generated Image Detection. *arXiv* **2025**, arXiv:2501.15253.
29. Kharel, A.; Paranjape, M.; Bera, A. DF-TransFusion: Multimodal Deepfake Detection via Lip-Audio Cross-Attention and Facial Self-Attention. *arXiv* **2023**, arXiv:2309.06511.
30. Kim, M.; Hong, J.; Ro, Y.M. Lip to Speech Synthesis with Visual Context Attentional GAN. *arXiv* **2022**, arXiv:2204.01726.
31. Zou, M.; Yu, B.; Zhan, Y.; Lyu, S.; Ma, K. Semantics-Oriented Multitask Learning for DeepFake Detection: A Joint Embedding Approach. *arXiv* **2025**, arXiv:2408.16305.
32. Datta, S.K.; Jia, S.; Lyu, S. Detecting Lip-Syncing Deepfakes: Vision Temporal Transformer for Analyzing Mouth Inconsistencies. *arXiv* **2025**, arXiv:2504.01470.
33. Ng, K.H.; Song, T.; Wu, Y.; Xia, Z. XLSR-MamBo: Scaling the Hybrid Mamba-Attention Backbone for Audio Deepfake Detection. *arXiv* **2026**, arXiv:2601.02944.
34. Xu, X.; Ren, Y.; Liu, L.; Wu, W.; Li, B.; Lu, C.; Wang, S.; Zhang, C. HoliAntiSpoof: Audio LLM for Holistic Speech Anti-Spoofing. *arXiv* **2026**, arXiv:2602.04535.
35. Wahab, H.; Ugail, H.; Jaleel, L. Ensemble-Based Deepfake Detection using State-of-the-Art Models with Robust Cross-Dataset Generalisation. *arXiv* **2025**, arXiv:2507.05996.
36. Jabra, M.B.; Cheikhrouhou, O.; BenAmor, A. Deepfake detection through ensemble learning. *Multimed. Tools Appl.* **2025**, *84*, 45417–45452. <https://doi.org/10.1007/s11042-025-20932-w>.
37. Rana, M.S.; Sung, A.H. DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection. In *Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*; IEEE: New York, NY, USA, 2020, pp. 70–75. <https://doi.org/10.1109/CSCloud-EdgeCom49738.2020.00021>.
38. Knorn, S.; Chen, Z.; Middleton, R.H. Overview: Collective Control of Multiagent Systems. *IEEE Trans. Control Netw. Syst.* **2016**, *3*, 334–347. <https://doi.org/10.1109/TCNS.2015.2468991>.
39. Wooldridge, M.; Jennings, N.R. Intelligent Agents: Theory and Practice. *Knowl. Eng. Rev.* **1995**, *10*, 115–152. <https://doi.org/10.1017/S0269888900008122>.

40. Joshi, S. Review of Autonomous and Collaborative Agentic AI and Multi-Agent Systems for Enterprise Applications. *Int. J. Innov. Res. Eng. Manag.* **2025**, *12*, 65–76. <https://doi.org/10.55524/ijirem.2025.12.3.9>.
41. Gutowska, A. What Is a Multiagent System? IBM, 2024. Available online: <https://www.ibm.com/think/topics/multiagent-system> (accessed on 28 September 2026).
42. Lakara, K.; Channing, G.; Rupprecht, C.; Sock, J.; Torr, P.; Collomosse, J.; Schroeder de Witt, C. MAD-Sherlock: Multi-Agent Debate for Visual Misinformation Detection. *arXiv* **2024**, arXiv:2410.20140.
43. Le, H.H.; Nguyen, V.S.T.; Dang, T.L.C.; Nguyen, V.T.K.; Nguyen, T.T.H.; Cao, H. Multimedia Verification Through Multi-Agent Deep Research Multimodal Large Language Models. *arXiv* **2025**, arXiv:2507.04410.
44. Huynh, D.L.; Seal, S.; AIA4S Consortium; Reid, D.; Carpenter, A.E.; Bender, A.; Spjuth, O. AI agents in drug discovery: Applications and case studies. *Drug Discov. Today* **2026**, *31*, 104650. <https://doi.org/10.1016/j.drudis.2026.104650>.
45. Li, J.; Grover, A.; Alpuerto, A.; Cao, Y.; Liu, X.Y. Orchestration Framework for Financial Agents: From Algorithmic Trading to Agentic Trading. *arXiv* **2025**, arXiv:2512.02227.
46. Delancy, S.L.; Liang, L. Investigating Deepfake Detection Using LIME. *J. Comput. Sci. Coll.* **2024**, *40*, 76.
47. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*; Association for Computing Machinery: New York, NY, USA, 2020, pp. 180–186. <https://doi.org/10.1145/3375627.3375830>.
48. Grinberg, P.; Kumar, A.; Koppiseti, S.; Bharaj, G. What Does an Audio Deepfake Detector Focus on? A Study in the Time Domain. In *Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2025, pp. 1–5. <https://doi.org/10.1109/ICASSP49660.2025.10887568>.
49. Mercea, O.B.; Hummel, T.; Koepke, A.S.; Akata, Z. Temporal and cross-modal attention for audio-visual zero-shot learning. *arXiv* **2022**, arXiv:2207.09966.
50. Tran, K.D. Explainable Manipulated Videos Detection Using Multimodal Large Language Models. In *Companion Proceedings of the ACM Web Conference 2025*; Association for Computing Machinery: New York, NY, USA, 2025, pp. 725–728. <https://doi.org/10.1145/3701716.3715283>.
51. Ye, Z.; Sun, M.; Gao, H.; Yu, C.; Shi, Y. MOAT: Evaluating LMMs for Capability Integration and Instruction Grounding. *arXiv* **2025**, arXiv:2503.09348.
52. Barrington, S.; Bohacek, M.; Farid, H. The DeepSpeak Dataset. *arXiv* **2024**, arXiv:2408.05366.
53. Croitoru, F.-A.; Hondru, V.; Popescu, M.; Ionescu, R.T.; Khan, F.S.; Shah, M. MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark. *arXiv* **2025**, arXiv:2505.11109.
54. Hou, Y.; Fu, H.; Chen, C.; Li, Z.; Zhang, H.; Zhao, J. PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset. *arXiv* **2024**, arXiv:2405.08838.
55. Chandra, N.A.; Lee, H.; Murtfeldt, R.; Qiu, L.; Karmakar, A.; Tanumihardja, E.; Farhat, K.; Caffee, B.; Lee, C.; Choi, J.; et al. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv* **2025**, arXiv:2503.02857.
56. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020, pp. 2886–2895. <https://doi.org/10.1109/CVPR42600.2020.00296>.
57. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv* **2020**, arXiv:2006.07397.
58. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020, pp. 3204–3213. <https://doi.org/10.1109/CVPR42600.2020.00327>.
59. Zhou, T.; Wang, W.; Liang, Z.; Shen, J. Face Forensics in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021, pp. 5774–5784. <https://doi.org/10.1109/CVPR46437.2021.00572>.
60. Kwon, P.; You, J.; Nam, G.; Park, S.; Chae, G. KoDF: A Large-scale Korean DeepFake Detection Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021, pp. 10724–10733. <https://doi.org/10.1109/ICCV48922.2021.01057>.
61. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Proceedings of the NeurIPS Datasets and Benchmarks Track*. *arXiv* **2021**, arXiv:2108.05080.
62. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023, pp. 9739–9748. <https://doi.org/10.1109/CVPR52729.2023.00939>.

63. Liu, J.; Wang, J.; Hou, S.; Ren, M.; Wu, H.; Ma, L.; Pei, R.; He, Z. Beyond Face Swapping: A Diffusion-Based Digital Human Benchmark for Multimodal Deepfake Detection. *arXiv* **2025**, arXiv:2505.16512.
64. Abbasi, M.; Váz, P.; Silva, J.; Martins, P. Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. *Appl. Sci.* **2025**, *15*, 1225. <https://doi.org/10.3390/app15031225>.
65. Ashok, V.; Joy, P.T. Deepfake Detection Using XceptionNet. In *Proceedings of the 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*; IEEE: New York, NY, USA, 2023, pp. 1–5. <https://doi.org/10.1109/RASSE60029.2023.10363477>.
66. Jbara, W.A.; Hussein, N.A.K.; Soud, J.H. Deepfake Detection in Video and Audio Clips: A Comprehensive Survey and Analysis. *Mesopotamian J. Cybersecur.* **2024**, *4*, 233–250. <https://doi.org/10.58496/mjcs/2024/025>.
67. Li, X.; Chen, P.Y.; Wei, W. Where are We in Audio Deepfake Detection? A Systematic Analysis over Generative and Detection Models. *ACM Trans. Internet Technol.* **2025**, *25*, 1–19. <https://doi.org/10.1145/3736765>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.