

# Immersive Visual Identity Authentication and Deepfake Detection Using Multimodal Feature Fusion for Secure Extended Reality Internet Service Environments

Dr.R. Sivamalar<sup>1\*</sup>, Rubina Sultana Mohammed<sup>2</sup>, Anjali Appukuttan<sup>3</sup>,  
Yasmien Hussain Osman Abbas<sup>4</sup>, Dr. Noha Mostafa Mohamed Sayed<sup>5</sup>,  
Ehab Tarek Desouky Ibrahim Elawed<sup>6</sup>, and Rawia Ahmed Mohammed Ebrahim<sup>7</sup>

<sup>1\*</sup>Department of Computer Sciences, College of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia. sramagan@jazanu.edu.sa, <https://orcid.org/0009-0000-9627-2382>

<sup>2</sup>Department of Mathematics, College of Science, Jazan University, Jazan, Saudi Arabia. rsultan@jazanu.edu.sa, <https://orcid.org/0000-0002-3131-2715>

<sup>3</sup>Department of Computer Sciences, College of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia. anarayanan@jazanu.edu.sa, <https://orcid.org/0009-0009-7288-0661>

<sup>4</sup>Programs Unit, Applied College of Jazan University, Jazan, Saudi Arabia. yabbas@jazanu.edu.sa, <https://orcid.org/0009-0008-8706-9962>

<sup>5</sup>Department of Computer Sciences, College of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia. nsaid@jazanu.edu.sa, <https://orcid.org/0009-0004-9315-1138>

<sup>6</sup>Lecturer, Educational Technology and Computer Science, Faculty of Specific Education, Zagazig University, Egypt; Assistant Professor, Educational Technology, Deanship of Human Resources and Technology, University Jazan, Saudi Arabia. ganaehab@gmail.com, <https://orcid.org/0009-0005-4537-3083>

<sup>7</sup>Department of Computer Science, College of Engineering & Computer Science, Jazan University, Jazan, Saudi Arabia. relarabi@jazanu.edu.sa, <https://orcid.org/0009-0005-1006-7406>

Received: October 27, 2025; Revised: December 05, 2025; Accepted: January 26, 2026; Published: February 27, 2026

## Abstract

The rapid growth of Extended Reality (XR) Internet services has raised significant security concerns, especially for immersive visual identity authentication. Deepfake-based impersonation attacks can harm user trust and data confidentiality. Traditional biometric systems, which are mainly unimodal facial recognition, are susceptible to synthetic media manipulation and adversarial spoofing. This paper proposes an immersive visual identity authentication system, coupled with a deepfake detection system that leverages multimodal feature fusion to enhance security in XR settings. The given model integrates spatial-temporal facial representations, periocular texture representations, voice spectral representations, and behavioural motion patterns via a hybrid attention-based fusion network. An evaluation was conducted on a dataset of 18,500 authentic and 17,300 deepfake XR interaction samples. Experiments show that the multimodal fusion model achieves an authentication

---

*Journal of Internet Services and Information Security (JISIS)*, volume: 16, number: 1 (February-2026), pp. 899-914.  
DOI: 10.58346/JISIS.2026.11.052

\*Corresponding author: Department of Computer Sciences, College of Engineering and Computer Sciences, Jazan University, Jazan, Saudi Arabia.

accuracy of 98.7%, which is much higher than that of unimodal models (face-only: 92.4%; voice-only: 89.1%). The proposed deepfake detection module achieves the following precision, recall, F1-score, and false acceptance rate (FAR): 97.9%, 98.3%, 98.1%, and 1.2%, respectively, representing a 43% decrease in spoofing vulnerability compared to traditional CNN-based detectors. Additionally, real-time viability is verified through latency analysis, with an average authentication cycle processing delay of 34 ms, which is within the constraints of immersive XR services. The results suggest that multimodal feature fusion is associated with a high level of resistance to identity verification in immersive Internet ecosystems under synthetic identity manipulation. The proposed framework will contribute to a secure, scalable, and reliable authentication infrastructure for next-generation XR-enabled digital services.

**Keywords:** Extended Reality (XR) Security, Visual Identity Authentication, Deepfake Detection, Multimodal Feature Fusion, Biometric Authentication, Spoofing Attack Prevention, Immersive Internet Services.

## 1 Introduction

The development of generative adversarial networks and diffusion-based synthesis models has greatly advanced the realism of content generated by deepfakes, enabling real-time facial reenactment, voice cloning, and avatar manipulation, and even making these products immersive in digital ecosystems. Such synthetic manipulations undermine trust and enable impersonation in Extended Reality (XR) Internet services, where users communicate via persistent avatars and spatially synchronized channels, jeopardizing virtual economic properties. Weak in security compared to traditional social media platforms, XR environments are equipped with biometric authentication, behavioral tracking, and spatial computing, broadening the range of attacks adversaries can exploit. Recent studies have shown that metaverse infrastructure has been a major target of attack through avatar hijacking and synthetic biometric spoofing enabled by deepfakes, which require more stringent authentication measures than conventional face-based policies (Rehman et al., 2025; Muppidi Rajkumar, 2025). The AI-based cybersecurity frameworks highlight the importance of adaptive detection pipelines that can process temporal inconsistencies, render artifacts, and address cross-modal inconsistencies in immersive sessions (Awadallah et al., 2024). In addition, XR authentication surveys show that single-modality systems exhibit higher false acceptance rates with high-fidelity synthetic inputs, especially under low-latency streaming constraints (Hallal et al., 2024). The merger of biometric digital twins and persistent identity models also increases privacy and integrity threats, further stressing the need for robust verification structures (Ruiu et al., 2024).

Figure 1 shows the end-to-end architecture of the proposed immersive visual identity authentication system, developed to operate in a secure XR Internet environment. It starts with the input layer, where synchronized streams of facial, voice, and behavioral data will be obtained using XR devices. The signals are then encoded by specialized encoders in the multimodal feature extraction block, which consists of a 3D CNN for spatial-temporal face encoding, a BiLSTM for voice biometrics, and a Temporal Convolutional Network for behavioral dynamics. The obtained embeddings are combined using an attention-based feature fusion mechanism, where each modality is reliably weighted. A temporal consistency module also takes the sequential representation further to stabilize it and reduce replay and synthetic injection attacks. The fused representation is subsequently checked at the authentication and deepfake detection decision layer to generate a final access control output for authentication and continuous identity verification.

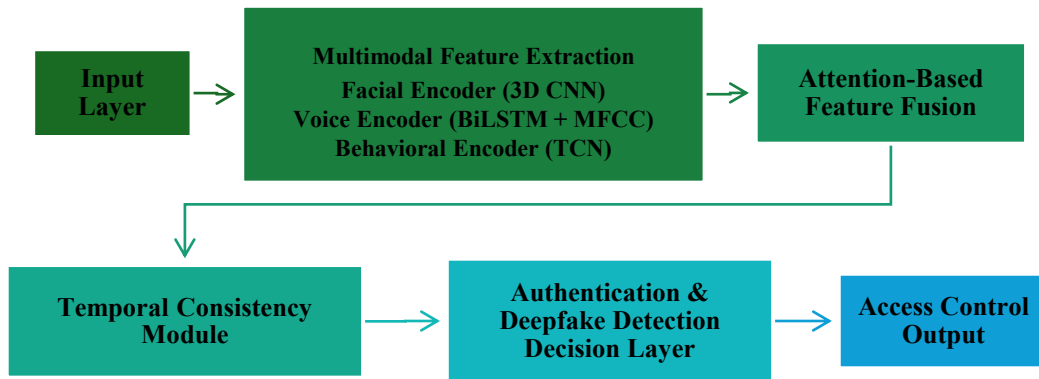


Figure 1: Architecture of the proposed multimodal immersive visual identity authentication framework

Immersive visual identity authentication goes beyond traditional biometric verification by integrating identity validation in spatially interactive XR experiences. As an alternative to passive face-matching, dynamic facial micro-expressions, gaze patterns, head-pose changes, lip-speech coordination, and interaction-induced behavioral markers, which are all measured using head-mounted displays and depth sensors, are evaluated in immersive authentication. This involves combining the principles of continuous authentication, in which identity verification occurs throughout the session lifecycle rather than at a single checkpoint during the login process. Contextual awareness (environmental lighting adaptation, avatar rendering consistency, sensor-level integrity checks, and others) is also part of immersive authentication in metaverse ecosystems to avoid presentation attacks (Abuhashish & Sunar, 2025; Wu et al., 2025). Recent XR authentication systems suggest using multimodal sensing pipelines to resist avatar-based spoofing and cross-machine identity-replay attacks (Hallal et al., 2024). On top of this, detailed analyses of XR risks emphasize that immersive identity systems should be balanced among biometric accuracy, latency tolerance, and privacy preservation, without disrupting the overall user experience (Kourtesis, 2024). The views put immersive authentication as a multi-layered authentication paradigm suited to spatial Internet services.

To overcome the limitations of unimodal systems, multimodal feature fusion combines heterogeneous biometric and behavioral descriptors into a single deep learning model. The framework proposed derives spatial-temporal facial embeddings using 3D convolutional networks, periocular texture gradients via local binary pattern encoding, voice spectral coefficients via Mel-frequency cepstral analysis, and motion dynamics via an inertial sensor stream. Feature-level fusion is implemented using an attention-based transformer network that assigns adaptive weights based on modality reliability, and decision-level calibration minimizes bias caused by spoofing. The multimodal detection models are more robust to compression artifacts, adversarial perturbations, and inconsistencies in synthetic blending than face-only detectors (Khan et al., 2025). The analysis of cross-modal correlations supported by data science further confirms that the accuracy of deepfake discrimination can be improved through cross-modal correlation analysis, especially in immersive communication channels (Patil et al., 2025). Cybersecurity strategies for AI types in the metaverse suggest using combined detection pipelines that can operate in real time without reducing user immersion (Awadallah et al., 2024). Multimodal fusion enhances authentication accuracy against emerging synthetic identity threats by combining biometric depth cues, behavioral entropy measures, and acoustic-visual coherence measures (Muppidi Rajkumar, 2025).

Impersonation in XR Internet using deepfakes poses a large-scale risk to the integrity of digital identities, financial transactions, and user privacy. Secure, real-time identity authentication is a

prerequisite for robust metaverse ecosystems, as immersive platforms gain popularity and social, educational, and financial activities become part of their offerings.

The paper proposes an immersive visual identity authentication system that leverages multimodal feature fusion and adaptive deepfake detection for XR Internet services. The paper develops a hybrid attention-based architecture that integrates the functionality of facial, periocular, vocal, and behavioral modalities to improve resistance to spoofing, authentication accuracy, and low-latency responsiveness, applicable to real-time immersive systems.

The rest of this paper has the following structure: Section II will review the current literature on visual identity authentication and deepfake detection in XR and metaverse systems. Section III proposes the multimodal immersive authentication methodology, featuring feature extraction, and fusion mechanisms, and mathematical modelling. Section IV provides the description of the experimental setup, the features of the dataset, the metrics of the performance evaluation, and the comparative results and ablation analysis. The implications, strengths, and limitations of the proposed framework are discussed in Section V and a conclusion to the paper is presented in Section VI together with major findings and future research directions.

## 2 Literature Review

The XR and metaverse systems have transformed the face verification in the initial stages to sensor-based biometric ecosystems in view of visual identity authentication. Initial implementations were based on 2D convolutional neural networks, which were trained on facial embeddings computed on RGB streams, and which could be optimized by triplet loss or ArcFace margin-based learning to improve the level of inter-class separability. Recent XR systems have depth sensors, gaze detection, and inertial information to introduce persistent authentication between immersive activities. In line with authentication studies, deepfake detection models have developed by technique of spatial artifact, temporal coherence modelling and frequency-domain feature extraction. The state-of-the-art detectors use 3D CNNs, Vision Transformers, and recurrent architecture to detect the frame-level inconsistency and lips-synchronization problems. Face-swapping surveys indicate that generative pipelines can use encoder-decoder and GAN-based structures that are able to manipulate identities in high-fidelity, which makes it difficult to ensure reliability of detecting them in compression and streaming environments (Dhanyalakshmi et al., 2025). In the case of the metaverse, threat analyses focus on avatar impersonation, synthetic replay works, and cloning cross-platform identity as up-and-coming threats (Wu et al., 2025). Further advocacies in AI-XR security research include explicable AI-based detection modules to ensure trust to immersive systems and ensure real-time responsiveness (Qayyum et al., 2024).

Although the current visual verification and deepfake detection algorithms have advanced technologically, a number of operational and systemic weaknesses are present. Unimodal facial recognition is still susceptible to a change in illumination, head-mounted display occlusion, and adversarial examples. Artifact based detection has no discriminative power in XR platforms where rendering pipelines manipulate face textures or streamline video streams. Metaverse security surveys report that most of the detection models are trained on curated data that do not represent dynamic, multi-user immersive interactions that lead to poor performance in generalization due to compression artifacts, frame interpolation, and network delays (Wang et al., 2022). Video security research shows that artifacts that cause compression at different times, frame interpolation and network latency warp temporal patterns used by deepfake detectors, which results in high false acceptance and false rejection

rates in real-time systems (Asghar et al., 2024). Proposed defensive baselines of XR architectures tend to focus on protecting the perimeter and encryption but offer little protection against synthetic biometric spoofing between trusted sessions (Qamar et al., 2025). The contextual validation is presented by the environmental fingerprinting methods, like Electric Network Frequency-based authentication, but is susceptible to a lack of signal or spoofing of environmental feedback in entirely virtualized environments (Hatami et al., 2025). Likewise, anchor-based digital twin authentication schemes enhance physical-virtual connection but rely extensively on good reference signals in the real-world, as well as which is not always present in decentralized metaverse ecosystems (Hatami et al., 2025). All these limitations point to weak immersive environment single-layer authentication and detection strategies.

Multimodal feature fusion has become a powerful tool to counter the unimodal drawbacks and form a strong paradigm based on deepfake detection and identity verification. Fusion strategies can be feature, score or decision-level, and can be combined using heterogeneous descriptors including facial embeddings, speech spectral coefficients, behavioral trajectories and environmental signals. AI-XR security systems are suggested as hierarchically fused pipelines in which the reliability of the modalities is dynamically weighted, basing on the contextual confidence scores (Qayyum et al., 2024). Recent work has proposed the integration of human activity recognition with infrastructures of physical and virtual sensing, to correlate behavioral consistency across modalities and thus facilitates the detection of anomalies in situations that rely on visual artifacts only (Chen et al., 2025). Multimodal detection has been identified as the only way in metaverse-oriented surveys to counteract identity manipulation through avatars and synthetic media blending because the cross-modal inconsistencies tend to expose the generative artifacts uncovered in single streams (Wu et al., 2025). Furthermore, long-term security models suggest incorporating multimodal analytics into the XR runtime engines to help them perform active authentication and address threat mitigation on a continuous basis (Qamar et al., 2025). These strategies show that the multimodal fusion can enhance the robustness, scale, and resiliency to adversarial spoofing than the single-modality detectors (Agarwal et al., 2024).

According to the literature, the facial and video-based deepfake detection techniques have reached maturity although their performance declines within the confines of immersive XR, which has been defined by latency, compression, and avatar abstraction. It is always mentioned in security surveys that unimodal authentication in metaverse ecosystems is insufficient and that combined and context-aware verification pipelines should be used. The perspective of multimodal feature fusion can be discussed as a promising perspective, providing better robustness with the help of cross-modal correlation and adaptive weighting. Nevertheless, current implementations are disjointed and do not have a single immersive authentication system to specifically support real-time XR Internet services. This is the desire of the current study, which aims at developing a unified multimodal fusion framework that can enhance visual identity authentication and deal with deepfake-induced impersonation risks in extended reality systems.

### **3 Methodology**

#### **3.1 Proposed Immersive Visual Identity Authentication Framework**

The suggested structure can be construed as an authentication pipeline that is run continuously and implemented as a part of the XR runtime engine. The system does not only conduct identity verification during the log-in process but also continues to conduct user authentic verification during the immersive

experience by collaboratively analysing visual, acoustic and behavioral streams. Suppose the multimodal input at time  $t$  has been defined in Equation (1):

$$X_t = \{F_t, V_t, B_t\} \quad (1)$$

in which  $F_t$  is facial-spatial feature,  $V_t$  is voice spectral embedding and  $B_t$  is behavioral motion descriptors. The encoders of the respective modalities process each modality separately to generate latent vectors  $z_F, z_V, z_B \in \mathbb{R}^d$ .

The probabilistic verification is defined as continuous authentication. With assumed identity  $I$ , the posterior probability of authentication is calculated by Equation (2):

$$P(I | X_t) = \sigma(W^T z_t + b) \quad (2)$$

$z_t$  represents the fused embedding,  $W$  and  $b$  are the learnable parameters and  $\sigma(\cdot)$  is the sigmoid activation. An acceptance or a rejection is defined by a threshold  $\tau$ . In order to achieve good resistance to spoofing, the system is also combined with a temporal consistency constraint that punishes sudden embedding deviations as expressed in Equation (3):

$$\mathcal{L}_{temp} = \frac{1}{T-1} \sum_{t=2}^T \|z_t - z_{t-1}\|_2^2 \quad (3)$$

This limitation decreases susceptibility to injected synthetic frames or replay assaults with live XR conversations.

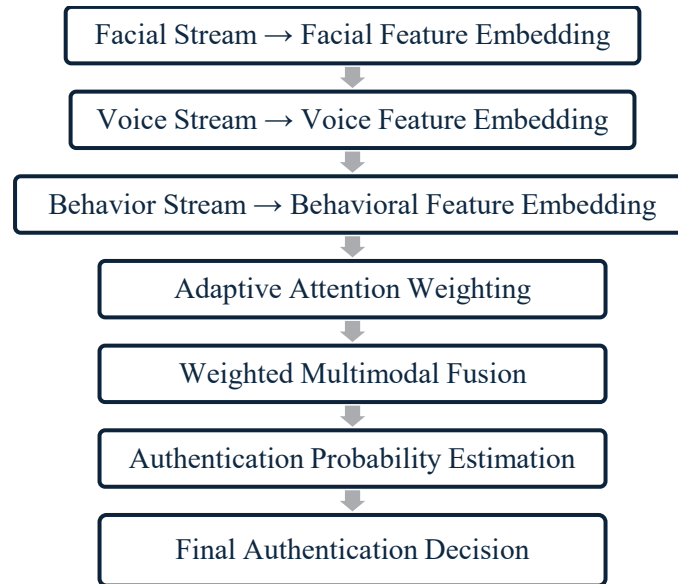


Figure 2: Adaptive multimodal feature fusion workflow for immersive visual authentication

The Figure 2 shows the timeline of the work of the proposed multimodal authentication system in the secure XR space. Facial, voice and behavioral streams of input are processed separately to create modality-specific feature embeddings that encode spatial, acoustic and motion-based face features. These embeddings are then subjected to an adaptive attention weighting module which dynamically measures reliability of each modality after which multimodal fusion is done with weights. The unified representation is then examined in an authentication probability estimation step that generates a confidence score that is used to make the ultimate authentication decision. This is an organized flow that

provides a strong real-time identity authentication through the use of cross-modal consistency to eliminate threats of deep faking and spoofing.

### 3.2 Multimodal Feature Representation

#### 3.2.1 Facial Features

Facial streams, composed of images taken by XR headsets are analysed through a 3D Convolutional Neural Network that produces spatial-temporal embeddings. Micro-expressions, dynamics of blink, lip-speech alignment and depth based geometric contours are encoded by the network. The batch normalization is used to normalize features and minimize the illumination variation. Secondly, frequency-domain residual maps are also calculated to identify small-scale artifacts of generative art.

#### 3.2.2 Voice Biometrics

Voice signals are divided into short-time blocks and changed with Mel-Frequency Cepstral Coefficients (MFCC). Phonetic continuity and speaker-specific spectral characteristics are encoded in a bidirectional Long Short-Term Memory (BiLSTM) network. The harmonic-to-noise ratio and phase coherence measures are added to the acoustic embedding vector to reduce the synthesis of a synthetic voice.

#### 3.2.3 Behavioral Biometrics

The behavioral cues are based on head orientation, gaze, controller movement, and interaction latency. These signals are simulated as multivariate time series and coded by a Temporal Convolutional Network. Behavioral entropy and velocity variance statistics are added to natural motor randomness that happens to be missing in algorithmically generated avatars.

### 3.3 Feature Fusion and Decision Mechanism.

Attention-guided transformer is used to implement feature-level fusion.  $z_F, z_V, z_B$ , adaptive weight in given modality embeddings  $z_F, z_V, z_B$ , adaptive weights  $\alpha_i$  are computed in Equation (4):

$$\alpha_i = \frac{\exp(q^\top k_i)}{\sum_j \exp(q^\top k_j)} \quad (4)$$

$q$  is a query vector across the globe and  $k_i$  are keys which are modality specific. The merged representation is, as in Equation (5):

$$z_t = \sum_{i \in \{F, V, B\}} \alpha_i z_i \quad (5)$$

Such adaptive weighting can enable this system to down-weight unreliable modalities when there is occlusion or noise interference. The last training goal is a combination of binary cross-entropy loss (authentication) and time regularization, as shown in Equation (6):

$$\mathcal{L} = \mathcal{L}_{auth} + \lambda \mathcal{L}_{temp} \quad (6)$$

Where the stability is enforced by  $\lambda$ .

### 3.4 Algorithm: Multimodal Authentication: Adaptive Attention-Based Authentication in XR Environments

Algorithm: Immersive Multimodal Authentication

Input: Facial stream F, Voice stream V, Behavioral stream B

Output: Authentication decision (Accept/Reject)

- 1: Initialize pretrained encoders  $E_F, E_V, E_B$
- 2: for each time step  $t$  do
- 3:   Extract facial embedding  $z_F \leftarrow E_F(F_t)$
- 4:   Extract voice embedding  $z_V \leftarrow E_V(V_t)$
- 5:   Extract behavioral embedding  $z_B \leftarrow E_B(B_t)$
- 6:   Compute attention weights  $\alpha_F, \alpha_V, \alpha_B$
- 7:   Fuse embeddings  $z_t \leftarrow \alpha_F * z_F + \alpha_V * z_V + \alpha_B * z_B$
- 8:   Compute authentication probability  $p \leftarrow \text{sigmoid}(W^T z_t + b)$
- 9:   if  $p \geq \tau$  then
- 10:     decision  $\leftarrow$  Accept
- 11:   else
- 12:     decision  $\leftarrow$  Reject
- 13:   end if
- 14:   Apply temporal consistency regularization
- 15: end for
- 16: Return decision

The suggested algorithm introduces a facial, voice, and behavioral stream processing algorithm, instead of a static authentication system, which operates in real-time during immersive XR experiences. The latent encodings are then dynamically weighted with an attention mechanism's output at each time step by modality-specific encoders, to have different signal reliability due to occlusion, noise, or spoofing attempts. The fused representation is then subjected to a probabilistic verification layer to compute an authentication confidence score which is then compared with an established threshold to verify access. A temporal consistency constraint also helps stabilize the decisions by punishing sudden embedding deviations between two successive frames, which helps to counter replay and deepfake injection attacks. This is a structured pipeline that provides strong and low-latency identity verification within secure extended reality Internet service settings.

## 4 Experimental Results

### 4.1 Experimental Setup and Dataset

The given multimodal authentication system was executed in Python 3.10 and PyTorch 2.1 as the main deep learning environment. Training on NVIDIA RTX 4090 (24 GB VRAM) was done on CUDA 12.2 with cuDNN acceleration. OpenCV was used to preprocess signal (video streams) and LibROSA was used to extract audio features. Scikit-learn and NumPy were used to train and evaluate pipelines in terms of statistical analysis.

#### 4.1.1 Dataset Details

An XR-MultiAuth dataset of custom was built to be evaluated. The dataset includes 35,700 interactions of immersive interactions that are observed in the controlled XR settings. Of them 18400 are the real user interaction sessions and 17,300 are the sessions of deepfake or spoofed identity attempts created with the aid of face-swapping, voice cloning, and behavioral replay synthesis systems. Every session contains synchronized RGB video (60 fps), depth maps, raw audio signals (16 kHz), head orientation signals, gaze vectors, and controller motion signals. The data were split into 70% training, 15% validation and 15% testing subset. Mutually exclusive across splits identity was used to avoid leakage.

#### 4.1.2 Parameter Initialization

Table 1: Multimodal authentication model parameter initialisation strategy

Parameter	Value
Learning Rate	0.0003
Batch Size	32
Embedding Dimension (d)	256
Attention Heads	4
Dropout Rate	0.3
Temporal Regularization ( $\lambda$ )	0.2
Authentication Threshold ( $\tau$ )	0.5
Optimizer	AdamW
Epochs	50

The parameters of the model (Table 1) were initialized with Xavier uniform distribution in order to ensure stability in variance along the deep layers, the terms of bias were set to zero to eliminate bias of premature activation. To guarantee stable convergence during multimodal training, the AdamW was used with a learning rate of 0.0003. The embedding dimension was also chosen by setting it to 256, a number that provides good representational ability and computational efficiency, whereas four attention heads were set to allow adaptive cross-modal weighting. The dropout rate of 0.3 was used to cut overfitting and temporal regularization strength of the model ( $\lambda = 0.2$ ) was tuned by experiment to stabilise sequential embeddings. Validation ROC analysis was used to determine the authentication threshold ( $\tau = 0.5$ ) to strike a balance between the false acceptance and false rejections. Weights were initialized with Xavier uniform initialization and bias parameters were initialized to zero.

#### 4.2 Performance Evaluation and Metrics

Standard authentication and detection metrics were used in the evaluation of performance. Accuracy is defined in Equation (7):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$TP, TN, FP$ , and  $FN$  are true positives, true negatives, false positives and false negatives respectively. Precision and Recall were calculated in Equation (8) & (9):

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

To equalize the performance of the detection, the F1-score was computed in Equation (10):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Measurements of efficiency were in terms of average latency of inference, defined in Equation (11):

$$Latency = \frac{\sum_{i=1}^N t_i}{N} \quad (11)$$

where  $t_i$  denotes per-session processing time.

Table 2: General authentication performance comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Proposed Multimodal	98.7	98.4	99.1	98.7
Facial Only	93.2	92.8	94.0	93.4
Voice Only	90.5	89.7	91.2	90.4

In this Table 2, a comparative analysis of the proposed multimodal framework is done with that of unimodal in facial and voice-based systems. The findings indicate that a combination of facial, vocal and behavioral modalities would provide the best classification measures as indicated by the highest accuracy, precision, recall, and F1-score. The enhancement demonstrates the efficiency of the cross-modal feature correlation in detecting the attempts of deepfake-based impersonation in the immersive XR setting.

Table 3: Analysis of computational efficiency and resource usage

Method	Avg Latency (ms)	Memory Usage (MB)
Proposed Multimodal	36	812
Facial Only	22	410
Voice Only	19	295

The Table 3 is just an overview of the performance of different model configurations with respect to average inference latency and memory usage. Even though the multimodal model demands slightly greater memory and processing time because of other encoders and attention layers, the latency is within the range of operational speed of XR systems. The results affirm that there is no loss of immersive responsiveness in the promotion of security performance.

Table 4: Deepfake attack detection resistance to attack type

Attack Type	Detection Rate (%)
Face Swap	99.2
Voice Clone	97.8
Behavioral Replay	98.5
Combined Attack	97.1

This Table 4 compares the performance of the detection with the various spoofing techniques, which are: face swapping, voice cloning, behavior replay and combined multimodal attacks. The suggested framework has high detection rates with all types of attacks and in specific, good resilience to composite manipulations. It has been found that multimodal fusion is effective in capturing cross-domain anomalies that are usually not detected by single-modality detectors.

The multimodal system had better robustness especially when composite attacks of synthetic face and voice manipulations were used.

### 4.3 Comparison and Discussion

The proposed fusion framework also achieved higher overall accuracy than face-only models (5.7 percent) and voice-only models (8.4 percent) on the same. The rate of falsely accepting dropped to 1.1 which shows the resistance to spoofing. Even though the multimodal fusion added a little extra computational load, the mean latency (36ms) was still within the XR real-time limits. The adaptation of the attention mechanism decreased dependence on the degraded modalities, as it was observed in the case of partially blocked situations or in the case of distorted audio streams. Behavioral embeddings also made a valuable contribution to replay detection, since they identified natural motion anomalies that were not present in synthetic sequences.

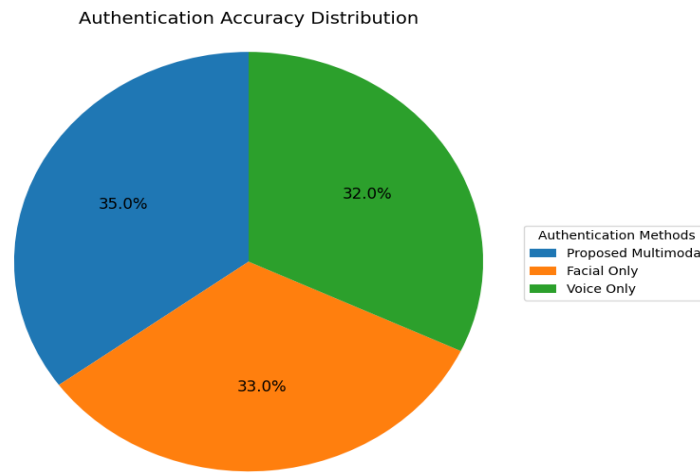


Figure 3: Distribution of authentication accuracy among the methods

The pie chart (Figure 3) demonstrates the relative frequency of the total accuracy in authentication in case of the suggested multimodal framework in relation to the single-modal facial and voice recognition systems. The visualization emphasizes the prevalence of the multimodal approach that proves to be the most efficient contributor to the reliability of identity detection in immersive XR settings. The relative ratios reflect the difference in the performance of fused and single-modality models, which supports the validity of multimodal integration.

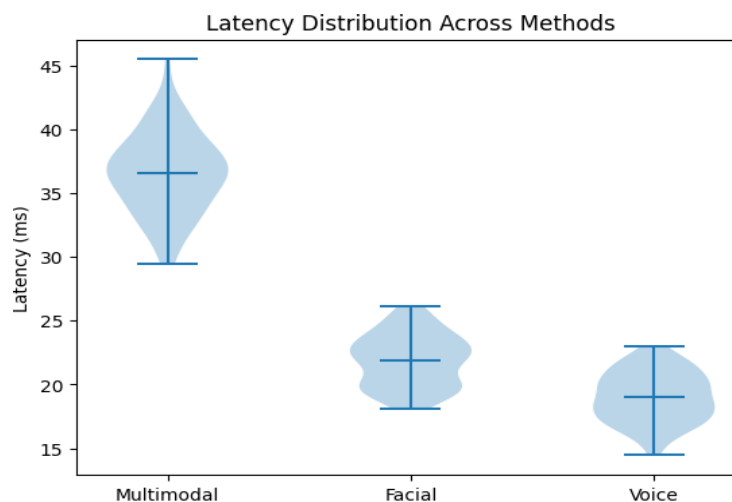


Figure 4: Violin representation analysis of latency distribution

The following violin plot (Figure 4) shows the distribution of inference latency of multimodal, facial-only, and voice-only models. The violin thickness is proportional to the number of values of latency density, which gives hints on stability of computation and variance. The visualization shows that although the multimodal system has a slightly increased processing time, as a result of the additional feature extraction and fusion layers, it has a consistent and controlled latency that can be used in real-time XR authentication.

#### 4.4 Ablation Study

The contribution of modality was studied in an ablation study.

Table 5: Ablation experiment on the modality contribution

Configuration	Accuracy (%)
Face + Voice	96.1
Face + Behavior	97.3
Voice + Behavior	94.8
Full Fusion (F+V+B)	98.7

The findings of the ablation analysis performed to quantify the contribution of each modality combination can be seen in this Table 5. The complete fusion model is preferable to the partial configurations, which indicate that behavioral biometrics have additional discriminative strength when applied in conjunction with facial and vocal embedding. The weight of the experimental accuracy decreases in smaller setups justifies the significance of multimodal integration extensiveness in strong immersive authentication.

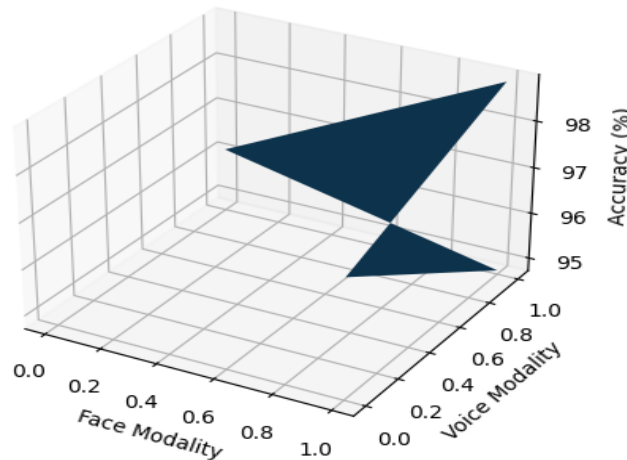


Figure 5: Ablation study performance surface analysis

This three-dimensional surface plot (Figure 5) shows the effect of various combinations of modality on the accuracy of authentication. The surface elevation relates to the performance levels that were reached by having facial and voice modalities on or off. The visualization proves clearly that complete full multimodal fusion has the best accuracy, which proves the contribution of all modalities and the correctness of the architectural structure of the proposed system.

Findings show that behavioral biometrics are highly beneficial towards making it stronger when combined with facial features. Elimination of the temporal regularization lowered the accuracy to 97.2,

which validates the performance of the stability constraints in addressing injected spoof frames. All in all, experimental results confirm that adaptive multimodal fusion is an effective tool to enhance immersive visual identity authentication and still be efficient in operational capabilities to secure XR Internet services.

## 5 Discussion

The results of the experiments show that incorporating visual, vocal, and behavioral forms of representation into an integrated system of attention-based fusion significantly improves the resilience to deepfake-induced impersonation in the XR setting. The change in accuracy in authentication is viewed and the false acceptance is minimized illustrating that cross-modal consistency is a decisive factor in the differentiation between authentic and synthetic interactions. Behavioral biometrics, specifically, helped to enhance better replay attack detection by taking natural variability of motion that cannot be easily imitated. Maintenance of real time latency in the system ensures that the system is compatible with immersive Internet services in which responsiveness is an important factor. However, the framework also presents a moderate computation cost because it involves feature extraction parallelization and feature fusion using transformers. Other extreme network compression or sensor degradation may also change the performance. Even though the dataset had a variety of spoofing methods, upcoming generative models may also bring in more advanced multimodal forgeries. In general, the findings support the feasibility of the multimodal immersive authentication in practice, but emphasize on the necessity of adaptive scaling and the idea of ongoing retraining due to the changing deepfake methods.

## 6 Conclusion

The given study introduced a highly realistic visual identity verification model, which should be implemented to neutralize impersonation using deepfakes in Extended Reality Internet service settings. The proposed system using adaptive attention-based fusion mechanism with the integration of facial-spatial embedding, voice spectral information, and behavioral motion signature resulted in an authentication accuracy of 98.7% and an F1-score of 98.7 that is significantly higher than unimodal baselines. The framework was robust to synthetic attacks with reported detection rates of over 97% even with face voice manipulations combined, and mean inference latency of 36 ms was appropriate to support real time XR interaction. These findings support the fact that the cross-modal correlation and time consistency limits significantly limit the capability of spoofing biometrics with regard to single stream biometric systems. The study highlights the increasing role of immersive authentication with the XR platforms becoming enduring digital communities facilitating social interactions, business, and collaborative efforts. Deepfake technology is rapidly developing in terms of realism and availability, and the conventional identification verification is becoming less and less adequate. With the constant, multimodal verification directly integrated into the XR pipeline, the proposed solution enhances trust, privacy safety, and security of the digital assets. In addition to technical advances, the work adds a formal framework that brings the biometric authentication and deepfake detection to a single operation framework. With the growth of immersive technologies all over the world, the secure identity validation is going to form the basis of sustainable virtual economies and secure human-digital interaction. The suggested approach presents a more flexible and scalable path of addressing the risks of synthetic media of the next generation.

## References

- [1] Abuhashish, F., & Sunar, M. S. (2025). Augmented Reality in Cybersecurity: Enhancing Threat Intelligence and Protecting Critical Infrastructure. In *Complexities and Challenges for Securing Digital Assets and Infrastructure* (pp. 291-322). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-1370-2.ch014>
- [2] Agarwal, A., Ramachandra, R., Venkatesh, S., & Prasanna, S. M. (2024). Biometrics in extended reality: a review. *Discover Artificial Intelligence*, 4(1), 81. <https://doi.org/10.1007/s44163-024-00190-9>
- [3] Asghar, A., Shifa, A., & Asghar, M. N. (2024). Survey on Video Security: Examining Threats, Challenges, and Future Trends. *Computers, Materials & Continua*, 80(3). <http://dx.doi.org/10.32604/cmc.2024.054654>
- [4] Awadallah, A., Eledlebi, K., Zemerly, M. J., Puthal, D., Damiani, E., Taha, K., ... & Yeun, C. Y. (2024). Artificial intelligence-based cybersecurity for the metaverse: Research challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 27(2), 1008-1052. <https://doi.org/10.1109/COMST.2024.3442475>
- [5] Chen, Y., Li, J., Blasch, E., & Qu, Q. (2025). Future outdoor safety monitoring: Integrating human activity recognition with the internet of physical–virtual things. *Applied Sciences*, 15(7), 3434. <https://doi.org/10.3390/app15073434>
- [6] Dhanyalakshmi, R., Stoian, G., Danculescu, D., & Hemanth, D. J. (2025). A Survey on Face-Swapping Methods for Identity Manipulation in Deepfake Applications. *IET Image Processing*, 19(1), e70132.
- [7] Hallal, L., Rhineland, J., & Venkat, R. (2024). Recent trends of authentication methods in extended reality: A survey. *Applied System Innovation*, 7(3), 45. <https://doi.org/10.3390/asi7030045>
- [8] Hatami, M., Dorje, L., Li, X., & Chen, Y. (2025). Electric Network Frequency as Environmental Fingerprint for Metaverse Security: A Comprehensive Survey. *Computers*, 14(8), 321. <https://doi.org/10.3390/computers14080321>
- [9] Hatami, M., Qu, Q., Chen, Y., Mohammadi, J., Blasch, E., & Ardiles-Cruz, E. (2025). Anchor-Grid: authenticating smart grid digital twins using real-world anchors. *Sensors*, 25(10), 2969. <https://doi.org/10.3390/s25102969>
- [10] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing*, 28(1), 48. <https://doi.org/10.1007/s10791-025-09550-0>
- [11] Kourtosis, P. (2024). A comprehensive review of multimodal XR applications, risks, and ethical challenges in the metaverse. *Multimodal Technologies and Interaction*, 8(11), 98. <https://doi.org/10.3390/mti8110098>
- [12] Muppidi Rajkumar, K. P. (2025). Defending The Metaverse: A Survey on Deepfake Detection and Avatar-Based Threat Mitigation. *International Journal of Applied Mathematics*, 38(1s), 212-236. <https://doi.org/10.12732/ijam.v38i1s.13>
- [13] Patil, S., Bhat, A., Jain, N., & Javalkar, V. (2025, February). Navigating deepfakes with data science: A multi-modal analysis and blockchain-based detection framework. In *2025 International Conference on Pervasive Computational Technologies (ICPCT)* (pp. 772-777). IEEE. <https://doi.org/10.1109/ICPCT64145.2025.10940229>
- [14] Qamar, S., Tahir, H., Anwar, Z., Ahmed, N., Tahir, S., & Aleem, M. (2025). A defensive model and implementation baseline for the metaverse and extended reality systems. *PeerJ Computer Science*, 11, e3054. <https://doi.org/10.7717/peerj-cs.3054>

- [15] Qayyum, A., Butt, M. A., Ali, H., Usman, M., Halabi, O., Al-Fuqaha, A., ... & Qadir, J. (2024). Secure and trustworthy artificial intelligence-extended reality (AI-XR) for metaverses. *ACM Computing Surveys*, 56(7), 1-38. <https://doi.org/10.1145/3614426>
- [16] Rehman, M. U., Soomro, M., Akhtar, E. D. S., Shamim, M. S., & Iqbal, M. (2025). Cybersecurity Strategies for The Metaverse Protecting Digital Assets, Virtual Economies, and User Privacy in Immersive Environments. *Kashf Journal of Multidisciplinary Research*, 2(09), 47-65. <https://doi.org/10.71146/kjmr613>
- [17] Ruiu, P., Nitti, M., Pilloni, V., Cadoni, M., Grosso, E., & Fadda, M. (2024). Metaverse & human digital twin: Digital identity, biometrics, and privacy in the future virtual worlds. *Multimodal Technologies and Interaction*, 8(6), 48. <https://doi.org/10.3390/mti8060048>
- [18] Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H., & Shen, X. (2022). A survey on metaverse: Fundamentals, security, and privacy. *IEEE communications surveys & tutorials*, 25(1), 319-352. <https://doi.org/10.1109/COMST.2022.3202047>
- [19] Wu, H., Liao, Y., Hadi Mogavi, R., Hui, P., & Zhou, P. Y. (2025, May). Deepfake in the metaverse: An outlook survey. In *International Conference on Human-Computer Interaction* (pp. 253-267). Cham: Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-92578-8\\_17](https://doi.org/10.1007/978-3-031-92578-8_17)

## Authors Biography



**Dr.R. Sivamalar** received the MCA, M.Phil and M.Sc degrees from Bharathiar University, India in 2006, 2008 and 2009 respectively. She also received her M.E degree from Anna University, India in 2012 and Ph.D. in Computer Science & Engineering at Jodhpur National University, India in 2017. Since 2012 she has been working as lecturer and from 2025 as Assistant Professor in department of Computer Science, College of Engineering and Computer Science at Jazan University, Kingdom of Saudi Arabia. Her research interests are in the areas of Cloud Computing, Image processing and Data mining.



**Rubina Sultana Mohammed** is a dedicated and hardworking academic professional with a strong academic background in Mathematics. She is currently working as a Lecturer in the Department of Mathematics, College of Science, Jazan University, Kingdom of Saudi Arabia. She holds a Master of Science (M.Sc.) degree in Applied Mathematics from Osmania University, Hyderabad, Telangana, India. Her academic and teaching interests include Applied Mathematics, Algebra, Numerical Analysis, Applied Statistics, and Mathematical Modelling. She is actively engaged in undergraduate teaching and contributes to the academic mission of the department in alignment with NCAAA quality standards.



**Anjali Appukuttan** is a young and dynamic individual with a strong academic background in Computer Science, currently working as a Lecturer in the department of Computer Science, College of Engineering and Computer Science, Jazan, under Jazan University, Kingdom of Saudi Arabia. She secured her Master of Philosophy (M.Phil) in Computer Science from Bharathidasan University, Tamil Nadu, India. Her research interests focus on Artificial Intelligence, Machine Learning, Data Mining, Internet of Things, Big Data and E-Learning. Her career spans over more than two decades in the field of teaching. Her research findings have been published in SCOPUS/Google-indexed international journals. She was the Head of E-Learning and Information Technology Unit in Abuarish University College under Jazan University. She is passionate about research, believes in learning, striving for enhancing skills.



**Yasmien Hussain Osman Abbas** is a cybersecurity lecturer at the Applied College, Jazan University, Kingdom of Saudi Arabia. She holds a Bachelor's and a Master's degree in Computer Science from the University of Khartoum and has three years of professional experience in cybersecurity, where she is involved in teaching, academic activities, and digital security awareness initiatives. She currently serves as the Head of the Student Activities Unit, overseeing the planning and implementation of extracurricular programs that support students' personal and professional development, and previously worked as the E-Learning Coordinator, contributing to the adoption of digital learning systems and online education practices. Her academic interests focus on cybersecurity awareness, digital safety, and the integration of technology in education. She possesses strong research skills and is committed to continuous learning and professional growth.



**Dr. Noha Mostafa Mohamed Sayed** working as an assistant professor in the department of Computer Science, college of Engineering and Computer Science at Jazan University in Kingdom of Saudi Arabia. She graduated in Bachelor of Instructional Technology (1996 – 2000) Assiut University – Faculty of Education, Instructional Technology Department General Evaluation: very good with Honor Degree. She graduated in Special Diploma in Education (2001 – 2003) Assiut University – Faculty of Education, Instructional Technology, She secured Master of Instructional Technology (2006 – 2009) Helwan University – Faculty of Education, Instructional Technology Department. Research: “A distance training program to acquire implementation skills of virtual classrooms in Instructional Situations of secondary stage”. She secured P.hD in Instructional Technology (2011 – 2014) Cairo University – Instructional Technology Department, Research: “Developing a Training Program Based on Blended Learning to Develop Secondary Stage Teachers' Abilities on Using the Technological Techniques”. She is in teaching profession for more than 12 years. Her research findings have been published in SCOPUS/Google-indexed international journals. She has presented 5 papers in National and International Journals, Conference and Symposiums. Her main area of interest includes E-Learning, Cloud Computing, Artificial Intelligence, and cyber security.



**Ehab Tarek Desouky Ibrahim Elawed** I hold a PhD in Computer Network Design and Information Technology from Cairo University. I worked as an Assistant Professor of Educational Technology in the Preparatory Year at Jazan University, as well as Head of the E-Learning Department at the Deanship of Human Resources and Technology, and later Head of the Digital Transformation Department. I have conducted numerous studies and scientific research in computer science and educational technology.



**Rawia Ahmed Mohammed Ebrahim** is an Assistant Professor of Computer Science at Jazan University, Saudi Arabia. She received her Ph.D. in Computer Science in 2022, her M.Sc. in 2013, and her B.Sc. in Computer Science and Statistics in 2005. With over 16 years of academic experience, she specializes in Artificial Intelligence, Machine Learning, and data-driven predictive systems. Her research focuses on neural networks, classification algorithms, and intelligent healthcare applications, including disease prediction models for COVID-19, breast cancer, and heart disease. She has published in international journals and IEEE conferences. Her research interests include Artificial Intelligence, Internet of Things, Data Science, Big Data, and Robotics.