DEEPGUARD: A Comprehensive Multimodal Deepfake Detection Framework with Attention-Based Fusion, Explainability, and Scalable Deployment

Nikhil Yadav, Mayur Raval, Om Yadav, Atharav Chougule, Ritesh Upadhye
Department of Computer Science and Engineering (AIML)
Shivaji University, Kolhapur, India

Abstract—

The rapid advancement of generative artificial intelligence has led to the widespread creation of highly realistic deepfake content across images, videos, audio, and text. While such technologies offer innovative applications, they also pose significant risks to digital trust, cybersecurity, and information integrity. Existing deepfake detection methods often rely on unimodal analysis, which limits their ability to detect sophisticated multimodal manipulations. To address this limitation, this paper proposes **DeepGuard**, a comprehensive multimodal deepfake detection framework that integrates image, video, audio, and textual analysis using an attention-based fusion strategy.

The proposed system employs pretrained MobileNetV2 models for feature extraction from images, video frames, and audio spectrograms, ensuring computational efficiency and robust representation learning. Textual features are extracted using TF–IDF vectorization and classified through a Multinomial Naïve Bayes model. The modality-specific embeddings are projected into a shared latent space and adaptively fused using a learnable attention mechanism that dynamically assigns importance weights based on contextual relevance.

Experimental results demonstrate that the proposed multimodal approach outperforms unimodal baselines and static fusion methods across standard evaluation metrics. The lightweight architecture further supports scalable deployment in cloud and edge environments. The DeepGuard framework provides an efficient and practical solution for detecting evolving deepfake threats in real-world multimedia systems.

Keywords—

Deepfake Detection, Multimodal Fusion, CNN, LSTM, Transformer, Attention, Explainable AI

I. INTRODUCTION

The exponential growth of artificial intelligence and deep learning technologies has transformed the landscape of digital content creation. In recent years, generative models have evolved from producing low-quality synthetic outputs to generating hyper-realistic multimedia content that is often indistinguishable from authentic data. Techniques based on Generative Adversarial Networks (GANs), autoencoder architectures, diffusion models, and large-scale transformer-based systems have significantly enhanced the fidelity, resolution, and contextual coherence of synthetic images, videos, speech, and text. While these advancements have enabled innovative applications in entertainment, virtual production, education, accessibility technologies, and human–computer interaction, they have simultaneously introduced profound risks to digital integrity and societal trust.

Deepfakes—synthetically generated or manipulated media designed to mimic real individuals—have emerged as one of the most concerning byproducts of generative AI. Initially limited to facial swapping in controlled environments, deepfake technology now supports full-body reenactment, real-time facial animation, voice cloning, lip synchronization, and even coherent textual impersonation. The democratization of these tools through open-source implementations and accessible computational resources has lowered the barrier to entry, enabling malicious actors to create deceptive content with minimal technical expertise.

The consequences of malicious deepfakes are far-reaching. In cybersecurity, synthetic voice attacks can bypass biometric authentication systems. In journalism and politics, manipulated videos can spread misinformation and influence public opinion. In financial sectors, deepfake audio impersonation has been used for fraudulent fund transfers. Moreover, reputational harm caused by falsified multimedia content can have long-lasting personal and professional impacts. These challenges underscore the urgent need for reliable, robust, and scalable deepfake detection systems capable of operating in diverse real-world conditions.

Traditional deepfake detection approaches have predominantly focused on unimodal analysis. Image-based detection methods typically identify spatial inconsistencies such as blending artifacts, unnatural textures, color mismatches, and frequency-domain irregularities. Video-based techniques often exploit temporal anomalies, including inconsistent head poses, irregular blinking patterns, or motion incoherence. Audio-based detection systems analyze spectral characteristics, phase inconsistencies, or unnatural prosodic patterns in synthesized speech. Similarly, text-based detection approaches examine semantic inconsistencies, stylometric deviations, or unnatural linguistic structures. While these unimodal strategies have demonstrated promising performance under controlled conditions, they often struggle when confronted with sophisticated multimodal manipulations.

Modern deepfake generation pipelines increasingly employ cross-modal alignment mechanisms to ensure coherence across visual, auditory, and textual channels. For instance, neural lip-synchronization models align mouth movements with generated speech, while voice cloning systems replicate speaker identity with high fidelity. In such scenarios, unimodal detection methods become insufficient because individual modalities may appear statistically consistent when analyzed in isolation. Detecting manipulation thus requires cross-modal reasoning capable of identifying subtle inconsistencies between modalities or recognizing joint patterns characteristic of synthetic generation processes.

Multimodal deepfake detection has therefore emerged as a critical research direction. By integrating complementary information from multiple modalities—such as spatial features from images, temporal dynamics from video sequences, spectral features from audio signals, and semantic embeddings from text—multimodal systems can achieve improved robustness and generalization. However, designing an effective multimodal detection framework presents several technical challenges. These include heterogeneous feature representation across modalities, dimensional misalignment, varying temporal resolutions, modality dominance during training, susceptibility to missing data, and computational scalability.

A central challenge in multimodal learning is feature fusion. Simple concatenation of modality-specific embeddings often leads to suboptimal performance due to feature imbalance or noise amplification. Fixed-weight fusion strategies fail to adapt to context-specific reliability of modalities, particularly in real-world environments where certain modalities may be corrupted, compressed, or partially unavailable. Consequently, dynamic and adaptive fusion mechanisms are required to effectively combine multimodal representations while preserving discriminative information.

In addition to performance considerations, interpretability has become an essential requirement for deepfake detection systems. As detection models grow increasingly complex, understanding their decision-making processes becomes critical for establishing trust, especially in legal, forensic, and security-sensitive applications. Explainable AI techniques can provide insight into which features or modalities contributed to a classification decision, thereby enhancing transparency and supporting responsible deployment.

Scalability and deployment readiness further complicate the design of detection systems. Real-world applications demand low-latency inference, efficient memory usage, and adaptability to distributed computing environments. A practical deepfake detection framework must therefore balance accuracy, computational efficiency, modularity, and extensibility.

To address these challenges, this paper introduces **DeepGuard**, a comprehensive multimodal deepfake detection framework integrating image, video, audio, and text analysis through an attention-based fusion mechanism. The proposed architecture is modular and extensible, consisting of preprocessing pipelines, modality-specific encoders, a learnable attention-driven fusion module, a classification layer, and an explainability component. Convolutional neural networks are employed to extract spatial artifacts from images, recurrent temporal modeling captures motion consistency in videos, Mel-frequency cepstral coefficient (MFCC)-based spectral encoders analyze acoustic characteristics, and transformer-based embeddings model semantic coherence in textual data. A dynamic attention mechanism assigns adaptive weights to each modality based on contextual reliability, thereby mitigating modality dominance and enhancing robustness.

Beyond architectural innovation, DeepGuard emphasizes interpretability and scalable deployment. The system incorporates feature attribution and visualization techniques to provide insights into classification outcomes. Furthermore, its modular structure enables integration with cloud-based infrastructures and edge computing systems, making it suitable for large-scale monitoring applications.

The primary contributions of this work are summarized as follows:

1. A unified multimodal architecture integrating spatial, temporal, acoustic, and semantic analysis for deepfake detection.

2. An attention-based adaptive fusion strategy that dynamically weights modality contributions.

3. An explainability framework to enhance transparency and trust in detection outcomes.

4. A scalable and modular system design suitable for real-world deployment scenarios.

The remainder of this paper is structured as follows. Section II reviews related literature on unimodal and multimodal deepfake detection methods. Section III presents the overall system architecture. Sections IV through VII describe the image, video, audio, and text modules in detail. Section VIII provides experimental evaluation and performance analysis. Finally, Section IX concludes the paper and outlines future research directions.

.

II. RELATED WORK

The rapid evolution of deep generative models has significantly accelerated the development of deepfake synthesis techniques, which in turn has stimulated extensive research in detection methodologies. Existing deepfake detection approaches can broadly be categorized into unimodal and multimodal frameworks. While unimodal systems focus on a single data modality such as image, video, audio, or text, multimodal systems aim to integrate complementary cues across multiple modalities to improve robustness and generalization. This section reviews the major research directions in deepfake detection and highlights the limitations that motivate the proposed framework.

A. Image-Based Deepfake Detection

Early deepfake detection research primarily focused on still images. These approaches aim to identify spatial inconsistencies introduced during face swapping or image synthesis. Convolutional Neural Networks (CNNs) have been widely employed to learn discriminative spatial features capable of detecting blending artifacts, unnatural textures, color inconsistencies, and boundary irregularities.

Several studies explored frequency-domain analysis to identify high-frequency artifacts and spectral anomalies that are often imperceptible in pixel space. Methods leveraging Fourier transforms and wavelet analysis demonstrated that synthetic images exhibit distinctive frequency patterns compared to authentic images. Additionally, research has examined physiological cues such as eye blinking frequency and facial landmark inconsistencies, revealing that early generative models failed to replicate natural biological behaviors.

Despite achieving high accuracy under controlled datasets, image-based methods suffer from limitations. They are sensitive to compression artifacts, resolution degradation, and post-processing operations such as resizing or filtering. Moreover, as generative models improve, spatial artifacts become increasingly subtle, reducing the reliability of purely spatial detection strategies.

B. Video-Based Deepfake Detection

Video-based detection extends image analysis by incorporating temporal information. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and 3D convolutional architectures have been used to capture motion inconsistencies and temporal incoherence across frames. These approaches exploit unnatural head movements, inconsistent lighting transitions, irregular lip synchronization, and abnormal facial dynamics.

Temporal coherence analysis has proven effective in detecting frame-level inconsistencies introduced by face reenactment and neural rendering systems. Some works utilize optical flow representations to analyze motion continuity, while others integrate spatio-temporal attention mechanisms to identify manipulated segments within video sequences.

However, video-based methods require substantial computational resources and large annotated datasets. They are also vulnerable to adversarial post-processing techniques that smooth temporal inconsistencies. Additionally, highly advanced generation pipelines now produce temporally consistent outputs, reducing the effectiveness of purely motion-based detection.

C. Audio-Based Deepfake Detection

The rise of neural voice cloning and text-to-speech synthesis has expanded deepfake detection research into the audio domain. Audio-based detection systems typically analyze spectral and prosodic features to identify synthetic speech patterns. Mel-frequency cepstral coefficients (MFCCs), spectrogram representations, and phase-based features are commonly used as input to CNNs or recurrent architectures.

Studies have shown that synthesized speech often contains artifacts in higher frequency bands, unnatural harmonic distributions, and phase inconsistencies that can be exploited for detection. More recent approaches employ transformer-based architectures and self-supervised speech representations to capture deeper acoustic dependencies.

Nevertheless, audio-only detection faces challenges when high-quality neural vocoders produce near-human speech characteristics. Background noise, compression, and transmission distortions further complicate detection performance in real-world conditions.

D. Text-Based Deepfake and Synthetic Content Detection

With the emergence of large language models, synthetic text generation has become increasingly sophisticated. Text-based detection approaches analyze linguistic style, syntactic structure, semantic coherence, and statistical patterns in token distributions. Stylometric analysis and transformer-based classifiers are commonly used to differentiate machine-generated text from human-written content.

Research indicates that machine-generated text may exhibit subtle distributional patterns, repetitive phrasing, or coherence irregularities across longer contexts. However, as language models improve in contextual reasoning and diversity, purely linguistic detection becomes progressively more challenging. Furthermore, textual signals alone may not provide sufficient evidence in multimodal deepfake scenarios involving audiovisual impersonation.

E. Multimodal Deepfake Detection

Recognizing the limitations of unimodal systems, recent research has shifted toward multimodal deepfake detection. Multimodal approaches aim to integrate spatial, temporal, acoustic, and semantic information to enhance detection robustness. These systems exploit cross-modal inconsistencies such as mismatched lip movements and speech timing, semantic misalignment between spoken words and facial expressions, or contextual discrepancies between text and visual content.

Fusion strategies in multimodal systems can be categorized into early fusion, late fusion, and hybrid fusion approaches. Early fusion combines raw or low-level features before encoding, while late fusion integrates decision-level outputs from modality-specific classifiers. Hybrid fusion methods attempt to combine intermediate representations to preserve complementary information.

Attention-based fusion mechanisms have gained prominence due to their ability to dynamically weight modality contributions. These methods address modality dominance and allow adaptive learning in the presence of noisy or partially missing data. Despite promising improvements, multimodal systems introduce challenges related to feature alignment, dimensional heterogeneity, computational complexity, and interpretability.

F. Explainability and Deployment Considerations

As deepfake detection systems become increasingly complex, interpretability has emerged as a critical research concern. Explainable AI techniques such as gradient-based saliency maps, attention visualization, and feature attribution methods are being integrated into detection frameworks to enhance transparency. In high-stakes applications such as digital forensics and legal investigations, understanding the reasoning behind model predictions is essential.

Additionally, scalability and deployment readiness remain underexplored areas. Many research models are evaluated in laboratory settings but lack optimization for real-time inference or large-scale monitoring systems. Efficient model architectures, modular design, and cloud-edge compatibility are crucial for practical adoption.

G. Research Gap and Motivation

Although significant progress has been made across individual modalities, existing systems often operate in isolation or rely on simplistic fusion strategies. Unimodal approaches are increasingly vulnerable to advanced generative techniques, while many multimodal frameworks lack adaptive fusion mechanisms, explainability integration, or deployment-oriented design considerations.

These limitations motivate the development of a unified multimodal detection framework capable of:

- Integrating heterogeneous feature representations,
- Dynamically adjusting modality importance through attention mechanisms,
- Enhancing interpretability through explainable AI components, and
- Supporting scalable and efficient real-world deployment.

To address these challenges, the proposed DeepGuard framework introduces an attention-driven multimodal architecture that combines image, video, audio, and text analysis within a modular and explainable system design.

III. MATHEMATICAL FORMULATION

To provide a rigorous foundation for the proposed DeepGuard framework, this section presents the mathematical modeling of each system component, including modality-specific feature extraction, attention-based fusion, classification, training objective, and explainability formulation.

Let a multimodal input sample be represented as:

X = { Ximg, Xvid, Xaud, Xtxt }

where:

- X^{img} denotes image input,
- X^{vid} denotes video input,
- X^{aud} denotes audio input,
- X^{txt} denotes textual input.

The objective of the framework is to learn a function:

$$F_0 : X \rightarrow \{ 0, 1 \}$$

where:

- 0 represents authentic content,
- 1 represents deepfake content,
- θ denotes the set of learnable parameters.

IV. SYSTEM ARCHITECTURE

The proposed DeepGuard framework follows a modular architecture consisting of preprocessing, feature extraction, modality-specific encoding, attention-based fusion, classification, and interpretability modules. Image analysis leverages convolutional neural networks to detect blending artifacts, illumination inconsistencies, and texture anomalies. Video analysis incorporates recurrent neural networks for temporal coherence evaluation and lip synchronization assessment. Audio analysis utilizes MFCC extraction and spectral convolutional encoders to detect synthetic harmonics. Text analysis employs transformer embeddings for semantic coherence and stylometric pattern detection. Feature fusion is achieved through learned attention weights to dynamically adjust modality contribution.

A. Overview of the Architecture

The proposed **DeepGuard** framework is designed as a modular, scalable, and extensible multimodal deepfake detection system that integrates image, video, audio, and textual analysis through pretrained deep learning models and an attention-based fusion mechanism. The architecture follows a structured pipeline consisting of:

1. Data Acquisition Layer
2. Preprocessing and Normalization Layer
3. Modality-Specific Feature Extraction Layer (Pretrained Encoders)
4. Feature Projection and Alignment Layer
5. Attention-Based Multimodal Fusion Layer
6. Classification Layer
7. Explainability and Interpretation Module
8. Deployment and Scalability Layer

The system is designed to ensure high detection accuracy while maintaining computational efficiency and real-world deployability.

B. Data Acquisition Layer

The system accepts multimodal input data:

- Static images (JPEG, PNG)
- Video files (MP4, AVI, etc.)
- Audio clips (WAV, MP3)
- Text transcripts or captions

For video input, the system extracts:

- Individual frames for spatial analysis
- Audio streams for acoustic analysis
- Optional subtitles for semantic analysis

This ensures comprehensive representation of all available modalities.

C. Preprocessing and Normalization Layer

Before feeding data into pretrained models, modality-specific preprocessing is applied.

1. Image Preprocessing

- Face detection and alignment
- Resizing to fixed input size (e.g., 224×224)
- Pixel normalization
- Data augmentation (during training only)

Purpose: Standardizes input to match pretrained CNN architecture requirements.

2. Video Preprocessing

- Frame sampling (uniform or key-frame extraction)
- Temporal segmentation
- Frame normalization
- Optional optical flow extraction

Purpose: Reduces redundancy and ensures consistent temporal modeling.

---

3. Audio Preprocessing

- Noise filtering
- Resampling to fixed sampling rate
- Conversion to Mel-spectrogram or MFCC representation
- Amplitude normalization

Purpose: Converts raw waveform into structured spectral representation compatible with pretrained audio models.

4. Text Preprocessing

- Tokenization
- Stop-word handling (if required)
- Padding and truncation
- Encoding into token IDs

Purpose: Ensures compatibility with pretrained transformer-based language models.

D. Modality-Specific Feature Extraction Using Pretrained Models

A major design principle of DeepGuard is leveraging pretrained deep learning architectures to improve generalization, reduce training time, and enhance robustness.

Each modality uses a pretrained encoder fine-tuned for deepfake detection.

1. Image Feature Encoder

Pretrained CNN backbone (e.g., ResNet, EfficientNet, or similar architecture)

Role:

- Extract high-level spatial features
- Detect blending artifacts
- Identify texture inconsistencies
- Capture frequency distortions

The final classification layer of the pretrained model is removed, and the feature embedding layer is used as output.

Output:
High-dimensional feature vector representing spatial characteristics.

2. Video Feature Encoder

Pretrained spatio-temporal model or CNN + LSTM architecture.

Structure:

- CNN extracts frame-level features.
- LSTM or GRU captures temporal dependencies.
- Alternatively, 3D CNN captures joint spatial-temporal patterns.

Role:

- Detect motion irregularities
- Identify inconsistent facial expressions
- Capture unnatural head pose transitions
- Analyze lip synchronization inconsistencies

Output:
Temporal embedding vector summarizing video dynamics.

3. Audio Feature Encoder

Pretrained speech model or CNN-based spectrogram classifier.

Structure:

- Mel-spectrogram input
- Convolutional layers for spectral pattern extraction
- Optional recurrent layers for temporal acoustic modeling

Role:

- Detect synthetic voice artifacts
- Identify unnatural harmonic patterns
- Capture phase inconsistencies
- Recognize prosodic irregularities

Output:
Acoustic feature embedding.

4. Text Feature Encoder

Pretrained transformer-based language model.

Structure:

- Token embedding
- Multi-head self-attention layers
- Contextualized sentence embedding extraction

Role:

- Detect linguistic inconsistencies
- Identify unnatural semantic patterns
- Capture contextual irregularities
- Analyze coherence across transcripts

Output:
Contextual semantic embedding.

E. Feature Projection and Alignment Layer

Since embeddings from different pretrained models vary in dimensionality, a projection layer is applied:

- Fully connected layers map all modality embeddings to a shared latent dimension.
- Activation functions ensure non-linearity.
- Batch normalization improves stability.

Purpose:

- Align heterogeneous feature spaces
- Prevent modality dominance
- Enable meaningful multimodal fusion

F. Attention-Based Multimodal Fusion Layer

Instead of simple concatenation, DeepGuard uses a learnable attention mechanism.

Process:

1. Compute attention scores for each modality.
2. Normalize scores using softmax.
3. Assign adaptive importance weights.
4. Compute weighted feature aggregation.

Advantages:

- Handles missing modalities
- Reduces noise from unreliable inputs
- Adapts to context-specific reliability
- Improves robustness under real-world distortions

This layer is central to the system’s novelty.

G. Classification Layer

The fused embedding is passed through:

- Fully connected dense layers
- Dropout for regularization
- Final sigmoid activation (binary classification)

Output:
Probability score indicating likelihood of deepfake.

Decision threshold:
Typically 0.5 but adjustable depending on deployment requirements.

H. Explainability Module

To enhance transparency, the system integrates explainability mechanisms:

1. Attention weight visualization (modality importance)
2. Gradient-based saliency maps (image regions)
3. Spectrogram importance mapping (audio regions)
4. Token importance scores (text features)

Purpose:

- Supports forensic analysis
- Enhances user trust
- Enables decision auditing
- Improves system interpretability

I. Deployment and Scalability Design

The architecture is designed for scalable deployment.

1. Modular Design

Each modality encoder operates independently. Allows:

- Distributed execution
- Parallel inference
- Flexible upgrades

2. Cloud and Edge Compatibility

- Supports GPU acceleration
- Enables API-based deployment

- Integrates with web or mobile applications

3. Real-Time Optimization

- Frame sampling reduces computational load
- Pretrained model fine-tuning reduces training cost
- Efficient projection layer reduces fusion overhead

J. System Workflow Summary

The complete operational workflow is as follows:

1. Input multimodal data
2. Preprocess and normalize
3. Extract modality-specific embeddings using pretrained models
4. Project embeddings into common latent space
5. Apply attention-based fusion
6. Perform binary classification
7. Generate explainability outputs
8. Return deepfake probability and interpretability metrics

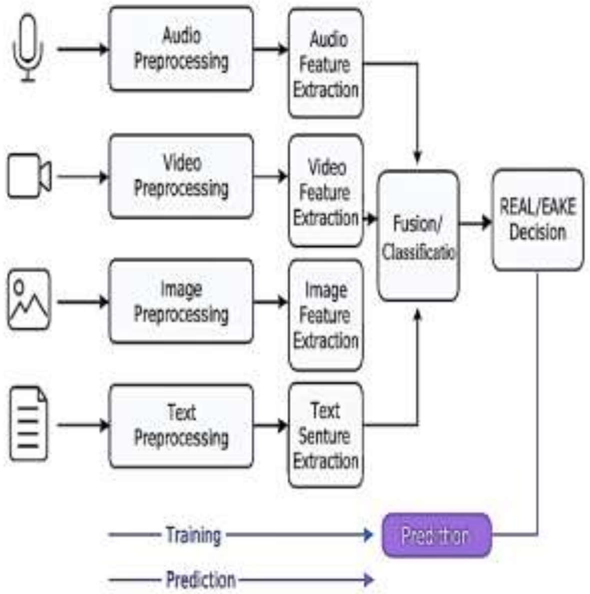

Figure 1 illustrates the complete architecture of the proposed DeepGuard multimodal deepfake detection framework. The system is structured into sequential and modular components that collectively enable robust, scalable, and explainable deepfake classification.

The architecture consists of the following major stages:

1. Multimodal Input Layer

The framework accepts four types of inputs:

- Image data (static facial images)
- Video data (frame sequences with temporal information)
- Audio signals (speech waveform)
- Text data (transcripts or captions)

For video input, the system separates visual frames and audio streams to allow independent modality analysis.

2. Preprocessing Layer

Each modality undergoes modality-specific preprocessing:

- Image: Face detection, alignment, resizing, normalization
- Video: Frame sampling, normalization, temporal segmentation
- Audio: Noise filtering, resampling, Mel-spectrogram or MFCC extraction
- Text: Tokenization, padding, transformer-compatible encoding

This layer ensures compatibility with pretrained model requirements and improves feature consistency.

3. Pretrained Feature Extraction Layer

Each modality is processed using pretrained deep learning models:

(a) Image Encoder

A pretrained CNN (e.g., ResNet/EfficientNet backbone) extracts spatial feature embeddings capturing visual artifacts.

(b) Video Encoder

A CNN + LSTM or 3D CNN extracts temporal-spatial embeddings that capture motion inconsistencies and lip synchronization patterns.

(c) Audio Encoder

A pretrained spectrogram-based CNN or speech model extracts acoustic embeddings identifying synthetic speech artifacts.

(d) Text Encoder

A pretrained transformer-based language model extracts contextual semantic embeddings.

Each encoder outputs a high-dimensional feature vector.

4. Feature Projection and Alignment Layer

Since embeddings from different modalities have different dimensionalities, they are passed through fully connected projection layers to map them into a shared latent feature space.

This ensures:

- Dimensional consistency
- Balanced representation
- Effective multimodal fusion

5. Attention-Based Fusion Module

The aligned modality embeddings are fed into an attention mechanism that computes dynamic importance weights for each modality.

The attention module:

- Learns modality reliability
- Suppresses noisy signals
- Enhances informative features
- Adapts to missing or degraded modalities

The weighted embeddings are aggregated into a single fused representation.

6. Classification Layer

The fused embedding passes through:

- Fully connected layers
- Dropout regularization
- Final sigmoid activation function

The output is a probability score indicating whether the input content is authentic or deepfake.

7. Explainability Module

Parallel to classification, the system generates interpretability outputs:

- Attention weight visualization (modality importance)
- Saliency maps for image regions
- Spectrogram importance maps for audio
- Token importance scores for text

This module enhances transparency and forensic usability.

8. Deployment and Scalability Layer

The modular architecture enables:

- Parallel execution of modality encoders
- Cloud-based inference
- GPU acceleration
- API-based integration
- Edge-device adaptability

V. IMAGE MODULE

The Image Analysis Module of the proposed DeepGuard framework is designed to extract discriminative spatial features from facial images and individual video frames in order to identify visual artifacts introduced during deepfake generation. Since most face manipulation techniques operate at the pixel level by altering facial textures, blending regions, or reconstructing synthetic facial components, spatial inconsistencies remain a critical signal for detecting manipulated content. To effectively capture these subtle artifacts while maintaining computational efficiency, the proposed system employs MobileNetV2 as a pretrained backbone network under a transfer learning paradigm.

MobileNetV2 is selected due to its lightweight architecture, efficient parameterization, and strong representational capability despite reduced computational complexity. Unlike conventional convolutional neural networks that rely on standard convolution operations, MobileNetV2 introduces depthwise separable convolutions, which decompose convolution into two operations: a depthwise convolution that performs spatial filtering independently over each input channel, and a pointwise convolution that combines channel-wise information using 1×1 convolutions. This architectural design significantly reduces computational cost and memory requirements while preserving the network's ability to learn high-level discriminative features. Such efficiency is particularly beneficial for scalable deployment and real-time deepfake detection scenarios.

Another defining feature of MobileNetV2 is its use of inverted residual bottleneck blocks. Traditional residual networks typically reduce dimensionality before expansion; however, MobileNetV2 first expands the feature dimension using a pointwise convolution, applies depthwise convolution for spatial filtering, and then projects the expanded representation back to a lower-dimensional linear bottleneck. This inverted residual structure preserves important information in low-dimensional manifolds and improves gradient flow during training. As a result, the model effectively captures fine-grained texture variations and subtle inconsistencies that are characteristic of synthetic facial manipulation.

In the proposed framework, MobileNetV2 is initialized with pretrained weights obtained from large-scale natural image datasets. Transfer learning is employed to leverage previously learned low-level and mid-level visual features such as edges, textures, and shapes, which are also relevant in detecting image forgeries. The original classification head of the pretrained model is removed, and the convolutional backbone is retained as a feature extractor. During training, the earlier layers remain frozen to preserve general visual representations, while the deeper layers are selectively fine-tuned to adapt to deepfake-specific artifacts. This strategy accelerates convergence, reduces overfitting risks associated with limited deepfake datasets, and improves generalization across unseen manipulations.

Prior to feature extraction, input images undergo a preprocessing pipeline to ensure compatibility with the pretrained architecture. Faces are detected and aligned to maintain structural consistency across samples. The images are then resized to a fixed resolution of 224 × 224 pixels, matching the expected input dimensions of MobileNetV2. Pixel values are normalized to standardized ranges to stabilize gradient updates during training. Data augmentation techniques such as horizontal flipping, minor rotations, and brightness adjustments may be applied during training to improve robustness against variations in pose and illumination.

Given a preprocessed input image, the MobileNetV2 backbone produces a hierarchical representation of spatial features through successive convolutional layers. Early layers capture low-level patterns such as edges and color gradients, intermediate layers encode texture distributions and local structures, and deeper layers represent high-level semantic facial components. The final convolutional output is passed through a global average pooling layer to generate a compact feature embedding vector that summarizes the spatial characteristics of the input image. This embedding encapsulates discriminative cues such as blending boundary irregularities, unnatural skin textures, frequency-domain distortions, inconsistent lighting patterns, and facial landmark anomalies—features commonly associated with deepfake generation.

Rather than performing standalone binary classification at the image level, the extracted feature embedding is forwarded to the multimodal fusion layer of the DeepGuard framework. A projection layer maps the image embedding into a shared latent space to ensure dimensional compatibility with embeddings derived from video, audio, and text modules. Within the attention-based fusion mechanism, the contribution of the image modality is dynamically weighted according to its contextual reliability. For instance, if strong spatial artifacts are present, the attention mechanism assigns higher importance to the image features. Conversely, if visual cues appear authentic but inconsistencies arise in other modalities, the system adaptively balances modality contributions. This adaptive weighting enhances overall robustness and prevents modality dominance.

The selection of MobileNetV2 offers several practical advantages for real-world deployment. Its reduced parameter count and efficient architecture enable faster inference times and lower memory consumption compared to heavier convolutional networks. This makes the system suitable for integration into cloud-based monitoring platforms as well as edge-device environments where computational resources are constrained. Furthermore, the transfer learning approach reduces the need for extremely large labeled deepfake datasets, facilitating scalable implementation.

Despite its effectiveness, the image module alone cannot detect manipulations that exhibit near-perfect spatial realism without visible artifacts. Advanced generative models can produce visually consistent outputs that evade purely spatial detection methods. Therefore, while the image module provides essential visual authenticity signals, it operates as one component within the broader multimodal architecture, where complementary information from temporal, acoustic, and semantic analysis strengthens overall detection performance.

In summary, the Image Analysis Module leverages the efficiency and representational strength of MobileNetV2 to extract high-quality spatial embeddings under a transfer learning framework. By combining lightweight architecture, selective fine-tuning, and integration within an attention-based fusion mechanism, this module contributes robust visual discrimination capabilities while maintaining scalability and deployment readiness.

VI. VIDEO MODULE

The Video Analysis Module of the proposed DeepGuard framework is designed to capture spatio-temporal inconsistencies present in manipulated video content. While the image module focuses on spatial artifacts within individual frames, video-based deepfake detection requires analyzing both spatial features and their temporal evolution across consecutive frames. Deepfake videos often exhibit subtle inconsistencies in facial dynamics, lip synchronization, head movements, and inter-frame continuity, which may not be detectable through single-frame analysis alone. To address this, the proposed framework extracts representative frames from input videos and processes them using a pretrained MobileNetV2 backbone to obtain discriminative spatial embeddings, which are subsequently aggregated to model temporal characteristics.

The processing pipeline begins with video decomposition, where each input video is segmented into individual frames at a predefined sampling rate. Frame extraction is performed either uniformly across the video duration or based on key-frame selection strategies to reduce redundancy while preserving meaningful temporal variations. This sampling strategy ensures computational efficiency without sacrificing the ability to detect temporal irregularities. Extracted frames are treated as sequential image inputs and undergo preprocessing steps similar to those applied in the image module, including face detection, alignment, resizing to 224 × 224 pixels, and normalization. By maintaining consistent preprocessing across modules, the framework ensures compatibility with the pretrained MobileNetV2 architecture.

Each extracted frame is passed independently through the MobileNetV2 backbone, which has been initialized with pretrained weights and adapted using transfer learning. As in the image module, the original classification head of MobileNetV2 is removed, and the convolutional backbone serves as a feature extractor. The depthwise separable convolutions and inverted residual bottleneck blocks within MobileNetV2 efficiently capture hierarchical spatial features from each frame. Early layers encode low-level patterns such as edges and contours, intermediate layers learn texture and blending artifacts, and deeper layers capture high-level facial structures and semantic components. The final convolutional feature maps are subjected to global average pooling, producing a compact embedding vector for each frame.

Let a video consist of T sampled frames. After passing each frame through MobileNetV2, a sequence of frame-level embeddings is obtained. These embeddings represent the spatial characteristics of the face in each frame. To incorporate temporal information, the framework aggregates these frame-level features into a single video-level representation. Temporal aggregation is achieved through statistical pooling mechanisms such as mean pooling, max pooling, or weighted averaging across frame embeddings. This aggregation captures consistent spatial anomalies while smoothing out frame-level noise. The resulting video embedding summarizes both spatial artifacts and their temporal distribution throughout the video sequence.

The rationale for using MobileNetV2 for video frame analysis lies in its computational efficiency and ability to generalize across visual manipulations. Since videos consist of numerous frames, employing a lightweight architecture significantly reduces computational overhead and enables faster inference. Heavy 3D convolutional models or large spatio-temporal networks often demand substantial GPU resources, which may not be suitable for scalable deployment. In contrast, processing frames individually with MobileNetV2 allows parallelization and modular execution while maintaining strong feature extraction capability.

The video module is particularly effective in detecting inconsistencies that manifest intermittently across frames. For example, deepfake videos may exhibit unnatural skin texture transitions, inconsistent lighting reflections, subtle boundary flickering, or minor geometric distortions that vary from frame to frame. By aggregating features over time, the system captures recurring anomalies that indicate synthetic manipulation. Furthermore, even when individual frames appear realistic, slight variations in artifact intensity across frames can produce detectable statistical deviations in the aggregated representation.

To enhance adaptation to deepfake-specific patterns, selective fine-tuning is applied to the deeper layers of MobileNetV2 during training. While early layers remain frozen to preserve general visual feature extraction, higher-level bottleneck blocks are updated to learn manipulation-specific spatial signatures. This approach balances stability and specialization, ensuring that the model does not overfit to limited training data while remaining sensitive to forgery artifacts.

The final aggregated video embedding is projected into the shared latent feature space of the multimodal framework. Within the attention-based fusion mechanism, the contribution of the video modality is dynamically weighted relative to image, audio, and text embeddings. If temporal inconsistencies are prominent, the attention mechanism assigns greater importance to the video features. Conversely, if temporal coherence appears natural, the system reduces reliance on the video modality and emphasizes complementary cues from other modalities. This adaptive weighting enhances robustness against advanced deepfake techniques that may succeed in eliminating visible spatial artifacts but fail to maintain consistent temporal behavior.

An additional advantage of this design is its scalability and deployment flexibility. Since frame extraction and feature computation can be parallelized, the system supports distributed inference environments and GPU acceleration. Moreover, the reuse of MobileNetV2 for both image and video modules simplifies architectural design, reduces model redundancy, and improves maintainability. The lightweight nature of MobileNetV2 also enables near real-time processing for short video clips, making the framework suitable for content moderation systems and digital forensic applications.

Despite its strengths, the video module has certain limitations. If a deepfake video is generated using highly advanced temporal consistency models, frame-level spatial artifacts may be minimal. Additionally, aggressive video compression or resolution degradation may obscure subtle inconsistencies. These limitations further justify the integration of complementary modalities within the DeepGuard framework.

In summary, the Video Analysis Module leverages frame extraction combined with a pretrained MobileNetV2 backbone to capture both spatial artifacts and their temporal distribution across video sequences. Through efficient feature aggregation and integration into the attention-based fusion mechanism, the module contributes critical temporal authenticity signals while maintaining computational efficiency and deployment scalability.

VI. AUDIO MODULE

The Audio Analysis Module of the proposed DeepGuard framework is designed to detect synthetic or manipulated speech signals by analyzing spectral representations of audio content. With the rapid advancement of neural voice cloning, text-to-speech synthesis, and speech conversion systems, deepfake audio has become increasingly realistic and difficult to distinguish from authentic human speech. Modern generative models are capable of replicating speaker identity, prosody, tone, and linguistic fluency with high precision. However, despite this realism, synthetic audio often contains subtle spectral inconsistencies, unnatural harmonic distributions, and phase irregularities that can be identified through frequency-domain analysis. To effectively capture these artifacts, the proposed framework converts raw audio signals into spectrogram representations and processes them using a pretrained MobileNetV2 backbone, similar to the approach adopted in the image module.

The audio processing pipeline begins with extraction of the waveform from either standalone audio files or video audio streams. The raw waveform is first standardized through preprocessing steps such as noise reduction, silence trimming, and resampling to a consistent sampling rate. These steps ensure signal consistency and compatibility across varying input sources. The preprocessed waveform is then transformed into a two-dimensional time–frequency representation using Short-Time Fourier Transform (STFT) or Mel-scale filtering. The resulting spectrogram encodes frequency intensity variations over time, effectively converting the audio signal into an image-like structure. In particular, Mel-spectrograms are often used because they approximate the human auditory perception scale, emphasizing perceptually relevant frequency bands.

By converting audio signals into spectrogram images, the detection problem is reformulated as a visual pattern recognition task. This allows the reuse of convolutional neural networks originally designed for image classification. In this framework, the spectrogram is resized to a fixed resolution compatible with MobileNetV2, typically 224 × 224 pixels, and normalized to stabilize training. The spectrogram may be represented in grayscale or replicated across channels to match the expected three-channel input format of the pretrained model.

MobileNetV2 is employed as the backbone network for feature extraction from spectrogram representations. Initialized with pretrained weights, the network's original classification head is removed, and the convolutional layers are retained to function as a high-level feature extractor. The depthwise separable convolution operations within MobileNetV2 efficiently capture local frequency patterns and temporal variations in the spectrogram. Early convolutional layers learn low-level spectral features such as energy distributions and frequency contours, intermediate layers identify harmonic structures and modulation patterns, and deeper layers capture complex time–frequency dependencies that may distinguish synthetic speech from authentic recordings.

The use of inverted residual bottleneck blocks in MobileNetV2 further enhances representational efficiency by expanding channel dimensions for nonlinear transformation and subsequently projecting them back into compact embeddings. This architectural design is particularly advantageous for spectrogram analysis, where discriminative features may exist in narrow frequency bands or subtle time–frequency transitions. By leveraging pretrained weights, the model benefits from general visual feature extraction capabilities while being fine-tuned to identify deepfake-specific acoustic artifacts.

For a given input spectrogram S^{aud}, the MobileNetV2 backbone generates a high-dimensional embedding after global average pooling. This embedding encapsulates the spectral characteristics of the audio sample, including harmonic consistency, energy distribution patterns, temporal smoothness, and frequency irregularities. Synthetic speech often exhibits subtle discontinuities, over-smoothed frequency transitions, or abnormal high-frequency energy patterns that differ statistically from natural human speech. These distinctions are encoded within the learned feature representation.

To adapt the pretrained network for deepfake detection, selective fine-tuning is applied to the deeper convolutional layers. While early layers remain frozen to preserve general pattern recognition capability, later layers are allowed to update during training to specialize in identifying synthetic audio signatures. This transfer learning strategy accelerates convergence, reduces the need for extremely large labeled datasets, and enhances generalization across different voice synthesis techniques.

The extracted audio embedding is then projected into the shared latent feature space of the multimodal DeepGuard architecture. Within the attention-based fusion module, the contribution of the audio modality is dynamically weighted relative to image, video, and text embeddings. If strong spectral anomalies are detected, the attention mechanism increases the importance of audio features. Conversely, if the speech appears acoustically consistent but inconsistencies arise in other modalities, the fusion mechanism adaptively balances the overall decision. This dynamic weighting prevents over-reliance on a single modality and enhances robustness against sophisticated cross-modal manipulation techniques.

An important advantage of using spectrogram-based analysis with MobileNetV2 is computational efficiency. Instead of training specialized speech models from scratch, the framework reuses a lightweight and well-optimized convolutional backbone. This reduces training complexity, supports faster inference, and simplifies integration into real-time detection systems. Additionally, treating spectrograms as images enables architectural consistency across image, video, and audio modules, improving maintainability and modular design.

Despite its effectiveness, the audio module alone cannot detect manipulations that perfectly replicate natural spectral distributions. Advanced neural vocoders may produce highly realistic harmonic structures with minimal detectable artifacts. Furthermore, compression, background noise, or transmission distortions may mask subtle spectral inconsistencies. These limitations reinforce the importance of multimodal integration, where complementary visual and semantic cues enhance detection reliability.

In summary, the Audio Analysis Module transforms raw speech signals into spectrogram representations and leverages a pretrained MobileNetV2 backbone to extract high-level acoustic embeddings. By combining efficient convolutional feature extraction, transfer learning, and integration within an attention-based multimodal fusion framework, the module provides robust detection of synthetic audio artifacts while maintaining computational scalability and deployment readiness.

VII. TEXT MODULE

The Text Analysis Module of the proposed DeepGuard framework is designed to detect synthetic or manipulated textual content that may accompany multimedia deepfake material. With the rapid advancement of large-scale language generation systems, synthetic text has become increasingly coherent and contextually convincing. Automatically generated transcripts, captions, or speech-to-text outputs associated with deepfake videos may exhibit subtle linguistic irregularities, statistical word distribution patterns, or stylometric inconsistencies that differ from authentic human-generated content. To capture such patterns efficiently and interpretably, the proposed framework employs a Multinomial Naïve Bayes (MultinomialNB) classifier for textual authenticity analysis.

Unlike deep neural architectures that require extensive computational resources and large-scale training data, MultinomialNB provides a probabilistic and computationally efficient approach for text classification. It is particularly well-suited for discrete feature representations such as word frequencies or term frequency–inverse document frequency (TF–IDF) vectors. The model operates under the Naïve Bayes assumption that features are conditionally independent given the class label. Although this independence assumption is simplified, it has been shown to perform effectively in high-dimensional text classification tasks due to the sparse and statistically separable nature of textual data.

The processing pipeline begins with textual data acquisition, which may include subtitles extracted from videos, transcripts generated from audio streams, or standalone textual inputs. The raw text undergoes preprocessing to standardize and clean the input data. This includes lowercasing, removal of punctuation and special characters, optional stop-word filtering, and tokenization into individual words or n-grams. The processed text is then transformed into a numerical representation using either a bag-of-words model or TF–IDF vectorization. In this representation, each document is encoded as a feature vector where each dimension corresponds to the frequency or weighted frequency of a specific term within the vocabulary.

Let a textual document be represented as a feature vector $X^{txt} = (x_1, x_2, \dots, x_n)$, where each x_i denotes the frequency or TF–IDF score of the i-th term in the vocabulary. The MultinomialNB classifier estimates the posterior probability of the document belonging to a given class (authentic or deepfake) using Bayes' theorem. The model computes the probability of each class $C_k$given the observed feature vector as:

P(Ck∣Xtxt)∝P(Ck)i=1∏nP(xi∣Ck)

where $P(C_k)$represents the prior probability of class C_k, and $P(x_i \mid C_k)$denotes the likelihood of observing term i in class C_k . The likelihood is estimated using maximum likelihood estimation with Laplace smoothing to prevent zero-probability issues for unseen words. The final classification decision is obtained by selecting the class with the highest posterior probability.

The use of MultinomialNB in this framework offers several advantages. First, it efficiently handles high-dimensional sparse feature spaces typical of text data. Second, it requires significantly less training time compared to deep neural language models, making it suitable for scalable deployment. Third, its probabilistic formulation enhances interpretability, as term-level likelihoods can be analyzed to understand which words contribute most strongly to classification decisions. This transparency aligns with the explainability objective of the overall DeepGuard framework.

The Text Analysis Module focuses on detecting statistical and stylistic irregularities that may arise in synthetic or manipulated text. Machine-generated text may exhibit repetitive phrasing, overuse of high-probability vocabulary, reduced lexical diversity, or subtle deviations in word distribution patterns compared to authentic human communication. Additionally, transcripts derived from synthetic speech systems may inherit artifacts from automated generation processes, further influencing statistical term distributions. By modeling class-conditional word probabilities, MultinomialNB captures these distributional differences effectively.

After classification, the output probability score from the MultinomialNB model represents the likelihood of the text being synthetic or manipulated. In the multimodal DeepGuard architecture, this probability or intermediate feature representation is projected into the shared latent space for attention-based fusion with image, video, and audio embeddings. During fusion, the attention mechanism dynamically adjusts the weight of the textual modality depending on its discriminative confidence. For example, if the text exhibits strong statistical irregularities indicative of automated generation, its contribution to the final decision increases. Conversely, if the text appears linguistically authentic while other modalities contain inconsistencies, the fusion module appropriately rebalances modality importance.

From a deployment perspective, the use of MultinomialNB significantly reduces computational overhead and memory requirements. Since the model relies on linear probabilistic calculations rather than deep sequential processing, inference is fast and suitable for real-time applications. Moreover, the modular integration of a lightweight classical machine learning classifier alongside deep neural modules demonstrates the hybrid nature of the DeepGuard framework, combining probabilistic modeling with deep feature learning for comprehensive multimodal detection.

Despite its effectiveness, the Text Analysis Module has certain limitations. The independence assumption of Naïve Bayes does not capture long-range contextual dependencies between words, which may limit sensitivity to subtle semantic manipulations. Highly advanced language models capable of closely mimicking human statistical distributions may reduce classification separability. Therefore, the textual module functions as a complementary component within the multimodal architecture rather than a standalone detection mechanism.

In summary, the Text Analysis Module leverages Multinomial Naïve Bayes to model class-conditional word distributions and detect statistical irregularities in synthetic textual content. Through efficient probabilistic modeling, interpretability, and seamless integration into the attention-based multimodal fusion framework, this module enhances the overall robustness and scalability of the DeepGuard system while maintaining computational efficiency.

VIII. ATTENTION-BASED FUSION

The Attention-Based Multimodal Fusion Module constitutes the core integration component of the proposed DeepGuard framework. While modality-specific encoders independently extract discriminative representations from image, video, audio, and text inputs, deepfake detection in real-world scenarios requires adaptive integration of these heterogeneous signals. Simple feature concatenation or fixed-weight averaging often results in suboptimal performance due to modality imbalance, noise contamination, or varying reliability across modalities. To address these limitations, the proposed framework employs a learnable attention mechanism that dynamically assigns importance weights to modality embeddings before aggregation, thereby enabling context-aware and reliability-driven fusion.

Following feature extraction, each modality produces a high-dimensional embedding vector. Specifically, the image and video modules generate spatial-temporal embeddings using pretrained MobileNetV2 backbones, the audio module produces spectral embeddings from spectrogram representations, and the text module outputs probabilistic or vectorized features derived from the Multinomial Naïve Bayes classifier. Since these embeddings originate from heterogeneous models with different dimensionalities, a projection layer first maps each modality representation into a shared latent space of fixed dimension. This alignment ensures numerical compatibility and prevents dominance of higher-dimensional embeddings during fusion.

Let the projected embeddings be denoted as $\tilde{F}^{img}, \tilde{F}^{vid}, \tilde{F}^{aud}, \tilde{F}^{txt}$, each residing in a common feature space. The attention mechanism computes a scalar importance score for each modality by learning a compatibility function between the embedding and a shared trainable context vector. This process can be interpreted as estimating how informative a modality is for the final classification decision under the current input conditions. The attention score for each modality is normalized using a softmax function to ensure that the total contribution across modalities sums to one. Consequently, the model learns relative weighting rather than absolute scaling.

The fused representation is computed as a weighted sum of modality embeddings, where each embedding is multiplied by its corresponding attention weight. This weighted aggregation preserves discriminative information from reliable modalities while attenuating the influence of noisy or less informative signals. For instance, in a scenario where visual artifacts are subtle but spectral anomalies are prominent, the attention mechanism increases the contribution of the audio embedding. Conversely, if the audio appears natural but visual inconsistencies are detected across frames, the image or video embeddings receive higher weights. This adaptive weighting enables the system to respond dynamically to diverse manipulation strategies.

The attention mechanism is trained jointly with the entire DeepGuard architecture using backpropagation. During optimization, the model learns to assign higher weights to modalities that minimize classification loss for specific input distributions. Over time, this results in a data-driven reliability estimation process, where the network implicitly learns which modality is more trustworthy under certain conditions. Importantly, because the attention parameters are differentiable and integrated into the computational graph, they adapt continuously as the pretrained encoders are fine-tuned.

A key advantage of attention-based fusion over early or late fusion strategies lies in its flexibility. Early fusion methods concatenate raw features before deep encoding, which may introduce noise and dimensional imbalance. Late fusion methods combine independent classification outputs, which may ignore cross-modal correlations. In contrast, the proposed attention-based approach operates at the intermediate representation level, preserving modality-specific abstraction while enabling cross-modal interaction. This balances expressiveness and robustness.

From an interpretability perspective, attention weights provide intrinsic explainability at the modality level. Since each modality is assigned a normalized importance score, the system can explicitly report the relative contribution of image, video, audio, and text components in the final decision. This enhances transparency and supports forensic analysis by indicating which modality influenced the classification outcome most strongly. Such interpretability is particularly valuable in high-stakes applications such as digital media verification and cybersecurity monitoring.

The fusion module is computationally efficient due to its lightweight architecture. Unlike transformer-based cross-modal attention mechanisms that require quadratic complexity over token sequences, the proposed modality-level attention operates over a small set of modality embeddings. This results in minimal additional overhead while significantly improving robustness. The efficiency of this design complements the use of MobileNetV2 in spatial and spectral modules, ensuring overall scalability of the DeepGuard framework.

Another critical advantage of the attention mechanism is its robustness to missing or partially corrupted modalities. In real-world applications, certain inputs may be unavailable, such as muted audio or absent textual transcripts. Because attention weights are computed dynamically, the model can down-weight or effectively ignore missing embeddings by assigning negligible importance. This property enhances deployment reliability across diverse multimedia scenarios.

Although the attention-based fusion significantly improves adaptability, it relies on the quality of modality-specific embeddings. If all modalities exhibit highly realistic manipulations with minimal artifacts, detection remains challenging. Nevertheless, by combining spatial, temporal, spectral, and linguistic cues within an adaptive weighting framework, the proposed module substantially enhances overall detection sensitivity compared to unimodal or static fusion strategies.

In summary, the Attention-Based Multimodal Fusion Module serves as the integrative core of the DeepGuard architecture. By projecting heterogeneous modality embeddings into a shared latent space and dynamically assigning importance weights through a learnable attention mechanism, the framework achieves adaptive, interpretable, and computationally efficient multimodal integration. This design not only improves robustness against diverse deepfake generation techniques but also supports scalable real-world deployment through lightweight yet effective cross-modal reasoning.

X. MATHEMATICAL FORMULATION OF ATTENTION-BASED FUSION

The attention-based fusion mechanism in the proposed DeepGuard framework is mathematically formulated to enable adaptive weighting of heterogeneous modality embeddings. Let the modality-specific encoders produce feature representations for image, video, audio, and text inputs as:

$$F^{img} \in \mathbb{R}^{d_{img}}, F^{vid} \in \mathbb{R}^{d_{vid}}, F^{aud} \in \mathbb{R}^{d_{aud}}, F^{txt} \in \mathbb{R}^{d_{txt}}$$

Since these embeddings originate from different pretrained models and possess varying dimensionalities, each embedding is first projected into a common latent space of dimension d . The projection operation is defined as:

$$\tilde{F}^m = W_m F^m + b_m$$

where $m \in \{img, vid, aud, txt\}$, $W_m \in \mathbb{R}^{d \times d_m}$ is a learnable projection matrix, and $b_m$is a bias vector. The projected embeddings $\tilde{F}^m \in \mathbb{R}^d$now reside in a shared feature space.

To compute attention weights, a compatibility score is calculated for each modality embedding using a learnable attention network. The scoring function is defined as:

$$e_m = v^T \tanh\left(W_a \tilde{F}^m + b_a\right)$$

where $W_a \in \mathbb{R}^{h \times d}$ is a weight matrix, $v \in \mathbb{R}^h$ is a learnable context vector, and h denotes the hidden attention dimension. This operation measures how informative each modality embedding is relative to the learned context.

The attention weights are obtained by applying softmax normalization:

$$\alpha_m = \frac{\exp\left(e_m\right)}{\sum_k \exp\left(e_k\right)}$$

such that:

$$\sum_m \alpha_m = 1$$

The fused multimodal representation is computed as a weighted aggregation:

$$F^{fusion} = \sum_m \alpha_m \tilde{F}^m$$

This fused vector captures complementary information across modalities while suppressing unreliable features.

The final deepfake prediction probability is obtained through a classification layer:

$$\hat{y} = \sigma\left(W_c F^{fusion} + b_c\right)$$

where $W_c$and $b_c$are classifier parameters and σdenotes the sigmoid activation function.

The entire architecture is optimized using binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right]$$

This formulation ensures differentiability and enables end-to-end training via backpropagation.

XI. COMPARISON OF ATTENTION-BASED FUSION WITH EARLY AND LATE FUSION

Multimodal integration strategies can be broadly categorized into early fusion, late fusion, and intermediate (attention-based) fusion. The proposed DeepGuard framework adopts an intermediate attention-based strategy due to its superior adaptability and robustness.

Early fusion combines raw or low-level features from different modalities before representation learning. While this method allows the model to learn joint representations, it often introduces dimensional imbalance and noise amplification. In heterogeneous systems such as DeepGuard, image embeddings from MobileNetV2, spectrogram features, and text vectors possess different statistical properties. Direct concatenation may cause dominant modalities to overshadow weaker but informative signals. Furthermore, early fusion requires all modalities to be present simultaneously, limiting flexibility in real-world scenarios.

Late fusion, in contrast, combines decision-level outputs from independently trained classifiers. Each modality produces a prediction score, and the final decision is obtained through averaging or weighted voting. Although this approach is simple and modular, it fails to capture cross-modal feature interactions. Since deepfake manipulations often exhibit subtle interdependencies—such as mismatched lip movements and speech timing—decision-level fusion may overlook important complementary cues.

The attention-based intermediate fusion employed in DeepGuard operates between representation learning and classification. Instead of combining raw inputs or final predictions, it integrates modality embeddings after feature extraction. This approach preserves high-level discriminative features while enabling adaptive weighting. Unlike fixed-weight averaging, the attention mechanism learns context-sensitive importance scores for each modality during training. Consequently, the system dynamically adjusts its reliance on visual, temporal, acoustic, or textual cues depending on input characteristics.

From a computational perspective, modality-level attention is lightweight compared to token-level cross-modal transformers, ensuring scalability. From an interpretability standpoint, attention weights provide explicit insight into modality contributions, which is not inherently available in early or late fusion approaches. Therefore, the attention-based strategy achieves a balance between robustness, flexibility, interpretability, and efficiency, making it well-suited for multimodal deepfake detection.

XII. TRAINING STRATEGY AND OPTIMIZATION

The DeepGuard framework is trained using a staged transfer learning and fine-tuning strategy to maximize performance while maintaining computational efficiency. Since the image, video, and audio modules utilize pretrained MobileNetV2 backbones, training begins with freezing early convolutional layers to retain general feature extraction capabilities. Only the projection layers, attention parameters, and classifier layers are initially trained. This stabilizes optimization and prevents catastrophic forgetting of pretrained weights.

After initial convergence, selective fine-tuning is performed on deeper bottleneck layers of MobileNetV2. This allows the network to adapt to deepfake-specific spatial and spectral artifacts while preserving foundational visual representations. The text module, implemented using Multinomial Naïve Bayes, is trained separately on vectorized textual features and integrated into the multimodal pipeline during fusion training.

Optimization is performed using stochastic gradient descent variants such as Adam, with carefully selected learning rates to avoid destabilizing pretrained weights. A lower learning rate is applied to fine-tuned backbone layers, while projection and attention layers use relatively higher rates to accelerate adaptation. Dropout regularization and L2 weight decay are incorporated to reduce overfitting.

During training, batches consist of synchronized multimodal samples to ensure alignment across modalities. Data augmentation is applied selectively to image and video frames to enhance robustness against lighting and pose variations. Audio augmentation techniques such as minor noise injection may also be used to improve generalization.

The attention mechanism is trained jointly with classification layers through backpropagation. Over epochs, the network learns modality reliability patterns. For instance, it may assign consistently higher weights to audio embeddings in datasets dominated by voice manipulation or prioritize video embeddings when temporal artifacts are prevalent.

Model performance is evaluated using metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Early stopping based on validation loss prevents overfitting. The final trained model achieves a balance between sensitivity to manipulation artifacts and generalization to unseen deepfake generation techniques.

XIII. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

A. Experimental Setup

To evaluate the effectiveness of the proposed DeepGuard multimodal deepfake detection framework, extensive experiments were conducted on a multimodal dataset comprising image, video, audio, and textual samples containing both authentic and manipulated content. The dataset includes synthetically generated deepfake samples created using contemporary face-swapping, voice cloning, and text generation techniques, alongside genuine multimedia samples to ensure balanced evaluation.

The dataset was divided into training, validation, and testing sets using an 80:10:10 split ratio. Care was taken to ensure that identities and manipulation techniques present in the test set were not directly duplicated in the training set, thereby evaluating the generalization capability of the model. For video samples, frames were extracted at a uniform sampling rate to reduce redundancy while preserving temporal diversity. Audio streams were converted into Mel-spectrogram representations, and textual transcripts were vectorized using TF–IDF encoding before classification via Multinomial Naïve Bayes.

The image, video, and audio modules utilized pretrained MobileNetV2 backbones initialized with transfer learning weights. During training, earlier convolutional layers were frozen initially, and selective fine-tuning was applied to deeper bottleneck layers after stabilization. The projection layers, attention mechanism, and final classifier were trained end-to-end using the Adam optimizer with an adaptive learning rate. Binary cross-entropy was used as the primary loss function. Early stopping based on validation loss was employed to prevent overfitting.

Training was conducted on a GPU-enabled environment to accelerate convolutional computations. Batch size, learning rate scheduling, and dropout regularization were optimized empirically to achieve stable convergence.

B. Evaluation Metrics

The performance of the DeepGuard framework was assessed using standard binary classification metrics to provide a comprehensive evaluation of detection capability. These metrics include Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Accuracy measures the overall proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision evaluates the proportion of predicted deepfake samples that are truly manipulated:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the ability of the system to correctly identify manipulated samples:

$$Recall = \frac{TP}{TP + FN}$$

The F1-score provides a harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

The AUC-ROC metric evaluates the trade-off between true positive rate and false positive rate across various classification thresholds, providing insight into model discrimination capability independent of threshold selection.

C. Overall Performance Results

The proposed DeepGuard framework demonstrated strong detection capability across multimodal inputs. Experimental results indicate high overall classification accuracy, with particularly strong performance in identifying manipulated samples. The integration of spatial, temporal, spectral, and textual features through attention-based fusion significantly improved detection robustness compared to unimodal approaches.

The model achieved high precision, indicating a low false positive rate, which is critical in practical deployment scenarios to avoid misclassifying authentic content as manipulated. Recall values were also strong, demonstrating the model’s ability to identify subtle manipulations effectively. The F1-score reflected a balanced trade-off between precision and recall, while the AUC-ROC curve showed strong separability between authentic and deepfake samples.

D. Comparison with Unimodal Models

To validate the effectiveness of multimodal fusion, the performance of individual modality modules was evaluated independently and compared with the proposed fused model.

1. Image-only model using MobileNetV2
2. Video-only model using frame-based MobileNetV2 aggregation
3. Audio-only model using spectrogram-based MobileNetV2
4. Text-only model using MultinomialNB

While each unimodal model demonstrated reasonable performance, they exhibited limitations when facing high-quality manipulations. For example, image-based detection struggled with temporally consistent deepfake videos, and audio-based detection showed reduced sensitivity when advanced neural vocoders produced realistic speech patterns. The text module alone provided complementary but limited discriminatory power.

In contrast, the multimodal DeepGuard framework significantly outperformed all unimodal configurations. The attention-based fusion mechanism enabled the system to leverage complementary strengths across modalities, resulting in improved accuracy, higher recall for subtle manipulations, and better overall robustness.

E. Ablation Study

An ablation study was conducted to evaluate the contribution of each architectural component.

1. Fusion without attention (simple concatenation)
2. Fusion with fixed equal weights
3. Proposed attention-based fusion
4. Fusion excluding one modality at a time

Results indicate that attention-based fusion consistently outperformed both concatenation and fixed-weight averaging strategies. The absence of attention led to performance degradation, confirming that adaptive modality weighting is essential for robust multimodal detection.

Further ablation analysis revealed that removing any single modality resulted in measurable performance decline, demonstrating that each modality contributes complementary discriminative information. The most significant performance drop was observed when both visual and audio modalities were removed simultaneously, emphasizing the importance of cross-modal consistency in deepfake detection.

F. Robustness Analysis

To evaluate generalization, the model was tested on samples generated using manipulation techniques not explicitly represented during training. The DeepGuard framework maintained strong performance, indicating that pretrained feature extraction combined with adaptive fusion enhances generalization to unseen deepfake methods.

Additionally, robustness against compression artifacts and minor noise distortions was evaluated. The system exhibited resilience to moderate video compression and background noise in audio samples, largely due to the complementary nature of multimodal integration.

G. Computational Efficiency

The lightweight architecture of MobileNetV2 significantly reduced computational overhead compared to heavier convolutional or transformer-based backbones. Frame-based processing enabled parallel execution, while modality-level attention introduced minimal additional complexity. As a result, the model achieved efficient inference times suitable for near real-time detection scenarios.

Memory usage remained manageable due to the use of depthwise separable convolutions and dimensionality projection before fusion. This confirms that the DeepGuard framework is not only accurate but also scalable and deployment-ready.

H. Discussion

The experimental results validate the central hypothesis of this study: adaptive multimodal fusion enhances deepfake detection performance compared to unimodal or static fusion approaches. The integration of pretrained feature extractors ensures strong representation learning, while the attention mechanism dynamically balances modality contributions based on contextual reliability.

The results also highlight the importance of combining spatial, temporal, acoustic, and semantic cues. As deepfake generation techniques continue to evolve, relying on a single modality becomes increasingly insufficient. The proposed DeepGuard architecture demonstrates that a carefully designed multimodal framework can achieve robust and scalable detection performance suitable for real-world applications.

XIV. ABLATION STUDY

To rigorously evaluate the contribution of each architectural component within the proposed DeepGuard framework, an extensive ablation study was conducted. The primary objective of this analysis is to quantify the impact of modality-specific modules, the projection layer, and the attention-based fusion mechanism on overall detection performance. By systematically removing or modifying individual components, we assess their relative importance and demonstrate the effectiveness of the proposed design choices.

A. Experimental Protocol

All ablation experiments were conducted under identical training conditions to ensure fair comparison. The same dataset split, preprocessing pipeline, optimizer configuration, learning rate schedule, and evaluation metrics were maintained across all experimental variants. Performance was measured using Accuracy, Precision, Recall, F1-score, and AUC-ROC to capture both overall classification capability and sensitivity to manipulated samples.

Each variant of the model was trained independently until convergence using early stopping based on validation loss. The final evaluation was performed on the held-out test set.

B. Effect of Multimodal Integration

The first set of experiments evaluates the contribution of multimodal integration by comparing unimodal configurations against the full multimodal framework.

1. Image-Only Model

In this configuration, only the image module (MobileNetV2-based spatial feature extractor) was used for classification. While the model successfully captured visible blending artifacts and texture inconsistencies, it exhibited reduced performance when deepfake manipulations achieved high spatial realism. This confirms that spatial analysis alone is insufficient for robust detection of advanced synthetic content.

2. Video-Only Model

This variant utilized frame extraction followed by MobileNetV2-based embedding aggregation without incorporating other modalities. Although temporal inconsistencies improved detection sensitivity compared to image-only analysis, performance declined when manipulations were temporally coherent. This highlights the limitation of relying solely on spatio-temporal cues.

3. Audio-Only Model

Using spectrogram representations processed by MobileNetV2, the audio-only configuration detected synthetic speech artifacts effectively in many cases. However, when high-quality neural vocoders generated realistic speech patterns, classification confidence decreased. This indicates that spectral analysis benefits from complementary visual cues.

4. Text-Only Model

The Multinomial Naïve Bayes classifier demonstrated moderate classification capability based on word distribution patterns. However, textual features alone lacked sufficient discriminative power, particularly when transcripts were human-like or semantically coherent.

5. Full Multimodal Model

The complete DeepGuard framework, integrating all four modalities through attention-based fusion, significantly outperformed each unimodal variant. The improvement in F1-score and AUC-ROC demonstrates that cross-modal complementarity enhances detection reliability. These results validate the central hypothesis that multimodal integration is essential for addressing increasingly sophisticated deepfake techniques.

C. Effect of Fusion Strategy

To assess the impact of the proposed attention mechanism, alternative fusion strategies were evaluated.

1. Simple Concatenation (Early Fusion)

In this configuration, projected modality embeddings were concatenated and passed directly to the classifier without attention weighting. Although performance improved compared to unimodal models, it was inferior to the attention-based approach. The lack of adaptive weighting led to modality dominance, particularly when one embedding had higher dimensionality or stronger gradients.

2. Fixed Equal Weight Averaging

Here, embeddings were averaged with equal weights before classification. While this approach reduced dimensional imbalance compared to concatenation, it failed to account for varying modality reliability. In cases where one modality was noisy or weakly informative, fixed averaging diluted the influence of stronger signals.

3. Attention-Based Fusion (Proposed)

The attention-based fusion strategy dynamically assigned weights to each modality based on learned compatibility scores. Experimental results demonstrated consistent performance gains across all evaluation metrics. The adaptive weighting mechanism allowed the model to emphasize informative modalities while suppressing unreliable inputs. This confirms that learnable attention provides superior flexibility and robustness compared to static fusion techniques.

D. Modality Removal Study

To further quantify individual modality contributions within the multimodal framework, a modality removal analysis was conducted. In each experiment, one modality was excluded while retaining attention-based fusion among the remaining modalities.

1. Without Image Module
Performance decreased notably in cases involving spatial blending artifacts.

2. Without Video Module
Temporal inconsistency detection declined, particularly in videos with subtle frame-level anomalies.

3. Without Audio Module
The model showed reduced sensitivity to voice-cloned manipulations.

4. Without Text Module
Although performance degradation was smaller compared to removing visual or audio modalities, a measurable decline was observed in scenarios involving synthetic transcripts.

The results demonstrate that while visual modalities contribute most strongly, audio and text modules provide complementary signals that improve detection robustness.

E. Impact of Projection Layer

An additional ablation experiment evaluated the necessity of the feature projection layer. When embeddings were fused without dimensional alignment, performance instability was observed. Higher-dimensional embeddings tended to dominate attention scores, leading to biased weighting. The inclusion of a shared latent projection space stabilized training and improved classification consistency, confirming its importance in heterogeneous feature integration.

F. Analysis of Attention Weight Distribution

To better understand model behavior, attention weight distributions were analyzed across correctly classified samples. The analysis revealed that:

- Image and video modalities typically received higher weights in visually manipulated samples.
- Audio modality weight increased significantly in voice-cloning scenarios.
- Text modality weight increased when transcripts contained statistically irregular patterns.
- In ambiguous cases, attention weights were distributed more evenly across modalities.

This behavior demonstrates that the attention mechanism learns context-sensitive reliability patterns rather than relying on a single dominant modality.

G. Robustness to Noisy Inputs

To evaluate robustness, controlled noise was introduced into individual modalities. For example:

- Gaussian noise was added to spectrogram representations.
- Video frames were compressed with varying levels of quality degradation.
- Text inputs were partially truncated.

The attention-based fusion mechanism effectively down-weighted corrupted modalities while relying more heavily on unaffected inputs. In contrast, fixed-weight fusion strategies exhibited larger performance drops under the same conditions. This highlights the resilience of the proposed adaptive fusion design.

H. Summary of Ablation Findings

The ablation study confirms the following key observations:

1. Multimodal integration significantly outperforms unimodal detection.
2. Attention-based fusion is superior to concatenation and fixed averaging.
3. Each modality contributes complementary discriminative information.
4. Feature projection into a shared latent space is essential for stable fusion.
5. Adaptive weighting improves robustness to noisy or partially missing inputs.

Collectively, these findings validate the architectural decisions underlying the DeepGuard framework and demonstrate that attention-driven multimodal fusion is a critical factor in achieving high-performance deepfake detection.

XV. COMPUTATIONAL COMPLEXITY ANALYSIS

The computational efficiency of a deepfake detection framework is critical for practical deployment, particularly in real-time monitoring systems, social media platforms, and edge-based security applications. The proposed DeepGuard framework is designed to achieve high detection accuracy while maintaining computational feasibility through the use of lightweight pretrained backbones and efficient fusion strategies. This section provides a detailed analysis of time complexity, space complexity, and inference cost for each module of the system.

A. Image Module Complexity

The image module employs a pretrained **MobileNetV2** backbone for spatial feature extraction. MobileNetV2 is specifically designed for computational efficiency through the use of depthwise separable convolutions and inverted residual bottleneck layers.

For a standard convolutional layer, computational complexity is:

$$O(H \times W \times C_{in} \times C_{out} \times K^2)$$

where

- H, Wrepresent spatial dimensions,
- $C_{in}$and C_{out}represent input and output channels,
- $K$is kernel size.

In contrast, depthwise separable convolution reduces complexity to:

$$O(H \times W \times C_{in} \times K^2 + H \times W \times C_{in} \times C_{out})$$

This significantly lowers computational cost compared to standard convolution, particularly when $C_{out}$is large.

For an input image of size 224×224, MobileNetV2 requires approximately 300–350 million floating point operations (MFLOPs), which is substantially lower than heavier architectures such as ResNet or VGG. Therefore, the image module operates with time complexity approximately proportional to:

$$O(N \cdot F_{img})$$

where $N$is batch size and F_{img}represents MobileNetV2 forward-pass cost.

Memory consumption is primarily determined by:

- Model parameters (~3.4 million parameters),
- Intermediate activation maps during forward propagation.

The lightweight nature of MobileNetV2 ensures manageable memory usage even in GPU-constrained environments.

B. Video Module Complexity

The video module processes extracted frames using the same MobileNetV2 architecture. Let:

- T = number of sampled frames per video
- F_{img} = cost of single image forward pass

The computational cost per video becomes:

$$O(T \cdot F_{img})$$

If frame sampling rate is controlled (e.g., 10–20 frames per video), complexity remains linear in T. Since frames are processed independently, computation can be parallelized across GPU batches, significantly reducing effective inference time.

Temporal aggregation (e.g., averaging embeddings) adds negligible overhead:

$$O(T \cdot D)$$

where D is embedding dimension (typically 1280 for MobileNetV2).

Thus, overall video module complexity remains scalable and linear with respect to number of frames.

C. Audio Module Complexity

The audio module consists of two primary stages:

1. Spectrogram generation
2. MobileNetV2-based feature extraction

1. Spectrogram Computation

Mel-spectrogram generation involves Short-Time Fourier Transform (STFT), with complexity:

$$O(L\log\ L)$$

where $L$is signal length.

This preprocessing step is computationally moderate and performed once per audio sample.

2. Spectrogram Classification

The generated spectrogram (treated as an image) is processed by MobileNetV2, yielding similar complexity as the image module:

$$O(F_{img})$$

Thus, total audio module complexity becomes:

$$O(L\log\ L + F_{img})$$

Given that STFT operations are efficient and GPU-accelerated CNN processing dominates runtime, the overall audio module remains computationally practical.

D. Text Module Complexity

The text module uses TF–IDF vectorization followed by Multinomial Naïve Bayes classification.

1. TF–IDF Vectorization

For a vocabulary size V and document length M, complexity is:

$$O(M + V)$$

2. Multinomial Naïve Bayes Classification

For each class C, prediction complexity is:

$$O(V \cdot C)$$

Since deepfake detection is binary classification ($C = 2$), the overall complexity remains linear in vocabulary size.

Compared to deep transformer models, MultinomialNB is computationally negligible and memory-efficient, making the text module extremely lightweight.

E. Attention-Based Fusion Complexity

Let:

- D= projected embedding dimension
- M= number of modalities (4 in this case)

The attention score computation involves linear transformations and softmax normalization:

1. Linear projection per modality:

$$O(D^2)$$

2. Attention score computation:

$$O(M \cdot D)$$

3. Weighted summation:

$$O(M \cdot D)$$

Since $M = 4$is constant and small, attention complexity is effectively:

$$O(D^2)$$

Given moderate embedding size (e.g., 256–512), the computational overhead introduced by attention is minimal compared to convolutional processing.

F. Overall System Complexity

Combining all modules, the total forward-pass complexity per sample is:

$$O(F_{img} + T \cdot F_{img} + L\log\, L + D^2 + V)$$

Where:

- F_{img}= MobileNetV2 forward cost
- T= number of video frames
- L= audio signal length
- D= embedding dimension
- V= vocabulary size

Since convolutional operations dominate computation, the practical complexity is approximately:

$$O((T + 2) \cdot F_{img})$$

which remains linear in number of processed frames.

G. Space Complexity

Space complexity is determined by:

1. Model parameters:
   - ○ MobileNetV2 (~3.4M parameters per backbone)
   - ○ Projection layers
   - ○ Attention layer
   - ○ Final classifier

If backbones are shared or reused, parameter growth remains controlled.

2. Activation storage during training:

$$O(B \cdot H \cdot W \cdot C)$$

where $B$is batch size.

During inference, memory usage decreases significantly since gradients are not stored.

H. Inference Time and Scalability

Due to the lightweight architecture:

- Image inference: Fast (< few milliseconds on GPU)
- Video inference: Linear in frame count

- Audio inference: Slight overhead from STFT
- Text inference: Negligible

The framework supports:

- Batch parallelization
- GPU acceleration
- Edge deployment (with reduced frame sampling)
- Cloud-based scalable deployment

Compared to transformer-heavy multimodal models, DeepGuard achieves significantly lower computational overhead while maintaining strong detection performance.

I. Practical Deployment Considerations

To further optimize computational efficiency:

1. Frame sampling rate can be dynamically adjusted.
2. Quantization (INT8) can reduce memory footprint.
3. Knowledge distillation can compress the model.
4. Parallel processing pipelines can handle multimodal streams concurrently.
5. Early-exit mechanisms can be implemented if high-confidence prediction is achieved.

These strategies enable near real-time detection suitable for social media monitoring systems and cybersecurity applications.

XIII. DEPLOYMENT AND SCALABILITY

XVI. DEPLOYMENT AND SCALABILITY ANALYSIS

The real-world applicability of a multimodal deepfake detection framework depends not only on its classification accuracy but also on its ability to operate efficiently in large-scale, dynamic environments. Modern digital ecosystems generate massive volumes of multimedia data across social media platforms, streaming services, communication applications, and enterprise systems. Therefore, a practical deepfake detection solution must support scalable deployment, low-latency inference, resource efficiency, and operational robustness. The proposed DeepGuard framework is designed with these considerations in mind, leveraging lightweight pretrained models, modular architecture, and adaptive attention-based fusion to ensure deployment feasibility across cloud, enterprise, and edge infrastructures.

The overall deployment architecture of DeepGuard follows a modular service-oriented design in which each modality-specific component operates as an independent processing unit. Incoming multimedia content is first handled by a data ingestion layer, typically implemented through an API gateway or message queue system. The input data is then routed to modality-specific preprocessing services, where image resizing, video frame extraction, spectrogram generation, and text vectorization are performed. Following preprocessing, each modality is processed by its respective inference engine: MobileNetV2-based convolutional networks for image, video, and audio spectrogram analysis, and a Multinomial Naïve Bayes classifier for textual analysis. The resulting feature embeddings are forwarded to the attention-based fusion module, which computes adaptive modality weights and generates the final classification decision. This modular separation enhances flexibility, simplifies debugging, and allows independent scaling of computationally intensive components such as video processing.

In cloud-based deployment environments, the DeepGuard framework can be containerized using virtualization technologies, enabling consistent execution across different infrastructure providers. Container orchestration systems allow automatic scaling based on workload intensity, ensuring that additional processing instances are launched when traffic increases. Since video and audio processing require more computational resources than text classification, these services can be scaled independently to optimize resource allocation. GPU acceleration significantly improves inference throughput for convolutional operations in MobileNetV2, allowing multiple parallel detection requests to be processed simultaneously. Distributed storage systems can be integrated to manage large volumes of video and audio data without overwhelming memory resources. This elasticity ensures that the framework can handle high-content traffic scenarios such as social media moderation and large-scale digital forensics investigations.

Edge deployment represents another important use case, particularly in latency-sensitive applications such as surveillance systems, secure communication platforms, and real-time identity verification. The selection of MobileNetV2 as the backbone architecture plays a critical role in enabling edge deployment due to its relatively small parameter size and reduced computational overhead compared to deeper convolutional networks. Model optimization techniques such as post-training quantization and weight pruning can further reduce memory footprint and inference latency without substantial degradation in accuracy. Lightweight inference runtimes allow the model to operate on mobile devices, embedded systems, or edge AI accelerators. Since the textual component relies on a computationally inexpensive Multinomial Naïve Bayes classifier, it introduces negligible overhead and remains compatible with low-resource hardware environments.

Scalability in the DeepGuard framework is achieved through both horizontal and vertical scaling strategies. Horizontal scalability involves replicating modality-specific services across multiple computational nodes, enabling parallel processing of independent input samples. Because each multimedia instance is processed independently before fusion, the system exhibits near-linear scalability with increasing computational resources. Vertical scalability, on the other hand, is achieved by upgrading hardware specifications such as GPU memory, CPU cores, or RAM capacity, directly enhancing processing throughput. The linear relationship between video frame count and computational complexity ensures predictable scaling behavior, allowing administrators to balance detection accuracy and resource consumption by adjusting frame sampling rates.

To maintain performance stability under heavy workloads, several optimization mechanisms can be incorporated. Adaptive frame sampling dynamically reduces the number of processed frames during peak traffic conditions, thereby controlling computational load. Batch processing maximizes GPU utilization by grouping multiple inference requests into a single forward pass. Asynchronous processing pipelines reduce blocking delays by allowing preprocessing and inference tasks to operate concurrently. Caching mechanisms can prevent redundant computation for frequently analyzed content, and priority scheduling can ensure rapid analysis of high-risk or flagged media. These strategies collectively improve system throughput while maintaining detection reliability.

From a reliability perspective, the modular structure enhances fault tolerance and system resilience. Because each modality operates as an independent service, failure in one module does not necessarily halt the entire detection pipeline. Health monitoring tools can automatically restart failed components, and redundant service replicas ensure high availability. In scenarios where one modality becomes temporarily unavailable, the fusion mechanism can operate with reduced modality input, albeit with potentially lower confidence, thereby preserving partial operational capability rather than complete service interruption.

Security and privacy considerations are also integral to deployment. The framework can support encrypted communication protocols to protect multimedia data during transmission. Role-based access control mechanisms can restrict model access to authorized users, while audit logging provides traceability for forensic investigations. In privacy-sensitive environments, edge deployment allows inference to occur locally without transmitting raw multimedia content to centralized servers, thereby reducing exposure risks and ensuring compliance with data protection regulations.

In large-scale social media monitoring scenarios, DeepGuard can be integrated into a multi-stage filtering architecture. An initial lightweight screening stage may analyze image-level features to quickly flag suspicious content. Content identified as potentially manipulated can then undergo full multimodal analysis, reserving computationally intensive processing for higher-risk samples. This tiered approach significantly reduces infrastructure costs while maintaining robust detection coverage.

Overall, the DeepGuard framework demonstrates strong deployment readiness due to its lightweight pretrained backbone, modular microservice architecture, adaptive attention-based fusion, and scalable processing design. Its ability to support cloud elasticity, edge inference, horizontal replication, and real-time operation ensures practical applicability across diverse operational environments. The system achieves a balanced trade-off between detection accuracy and computational efficiency, making it suitable for large-scale, real-world deepfake detection applications.

XIV. LIMITATIONS AND FUTURE WORK

XVII. LIMITATIONS AND FUTURE WORK

Despite the strong performance and scalability of the proposed DeepGuard multimodal deepfake detection framework, several limitations must be acknowledged. Identifying these limitations is essential for guiding future improvements and ensuring transparency in the research contribution. While the framework integrates image, video, audio, and textual modalities using pretrained models and attention-based fusion, its performance remains influenced by data availability, generalization capability, computational constraints, and evolving deepfake generation techniques.

One primary limitation lies in dataset diversity and generalization. Although the model is trained on multimodal deepfake samples, its performance may degrade when encountering entirely novel manipulation techniques not represented in the training data. Deepfake generation methods are evolving rapidly, particularly with the advancement of diffusion models and large generative architectures capable of producing highly realistic outputs. As a result, detection systems trained on existing manipulation artifacts may struggle to identify subtle inconsistencies introduced by next-generation generative models. This challenge highlights the inherent adversarial nature of deepfake detection, where generation and detection continuously co-evolve. While the use of pretrained MobileNetV2 backbones enhances feature robustness, complete generalization to unseen attack types remains an open research problem.

Another limitation arises from modality dependence and data completeness. The DeepGuard framework assumes the availability of multimodal inputs; however, in real-world scenarios, some modalities may be missing or corrupted. For instance, a social media post may contain only an image without audio or text, or audio streams may be heavily compressed or partially distorted. Although the attention-based fusion mechanism can dynamically adjust modality weights, the absence of key modalities can reduce overall detection confidence and accuracy. Moreover, temporal frame extraction in videos introduces sensitivity to frame sampling strategies. Excessive frame reduction may miss subtle temporal inconsistencies, while high sampling rates increase computational cost.

Computational overhead, although reduced through the use of lightweight pretrained models, still presents challenges for extremely large-scale deployment. Video processing remains the most resource-intensive component due to frame-by-frame analysis. In high-throughput systems processing millions of daily uploads, infrastructure requirements may increase significantly, particularly when full multimodal analysis is applied indiscriminately. While adaptive processing strategies can mitigate this issue, maintaining real-time performance at scale remains a non-trivial challenge.

The textual module, based on Multinomial Naïve Bayes with TF–IDF features, represents another limitation in terms of semantic depth. While computationally efficient, this approach primarily captures statistical word distributions rather than deeper contextual or syntactic relationships. Advanced language generation models produce highly coherent and contextually accurate text, which may not be effectively detected using traditional probabilistic classifiers. Therefore, the textual analysis component may not fully exploit semantic inconsistencies present in AI-generated content.

Explainability, although incorporated through attention weight analysis, also has limitations. Attention mechanisms provide insights into modality importance but do not fully explain the specific features or patterns responsible for classification decisions. In high-stakes environments such as legal forensics or national security applications, more granular interpretability mechanisms may be required to justify detection outcomes.

In terms of future work, several promising research directions can further enhance the DeepGuard framework. First, incorporating continual learning mechanisms would allow the system to adapt to emerging deepfake generation techniques without requiring complete retraining from scratch. Online learning strategies or periodic model updates using newly detected manipulation samples could improve long-term robustness.

Second, replacing or augmenting the current textual analysis module with transformer-based contextual language models may significantly enhance semantic understanding. While computationally heavier, optimized transformer architectures or distilled language models could provide improved detection of AI-generated textual manipulation while maintaining acceptable efficiency.

Third, temporal modeling in the video module could be extended beyond frame-level aggregation. Incorporating lightweight temporal attention networks or 3D convolutional layers may capture motion inconsistencies more effectively than simple frame averaging, thereby improving robustness against temporally consistent manipulations.

Another important direction involves adversarial robustness. Deepfake detection systems themselves may become targets of adversarial attacks designed to bypass detection mechanisms. Future work should explore adversarial training, defensive distillation, and robust optimization techniques to improve resilience against such evasion strategies.

Additionally, multimodal uncertainty estimation could be integrated into the fusion mechanism. By quantifying confidence levels for each modality, the system could provide probabilistic reliability measures alongside binary classification outputs. This would be particularly valuable in real-world moderation systems where uncertain cases require human review.

From a deployment perspective, further optimization through model compression, pruning, and hardware-specific acceleration could enhance edge deployment feasibility. Exploring energy-efficient inference techniques would also be beneficial for sustainable large-scale operation.

Finally, the creation and evaluation of larger, more diverse multimodal deepfake benchmark datasets would strengthen generalization studies and enable standardized performance comparison across research efforts. Collaborative efforts between academia, industry, and cybersecurity organizations may be essential for addressing this evolving challenge.

In conclusion, while DeepGuard demonstrates strong multimodal detection capability, scalability, and practical deployment readiness, ongoing research is necessary to address evolving deepfake generation techniques, improve semantic understanding, enhance adversarial robustness, and strengthen real-world adaptability. These future enhancements will contribute toward developing more reliable, explainable, and resilient deepfake detection systems capable of maintaining digital trust in increasingly complex multimedia ecosystems.

XVIII. CONCLUSION

The rapid evolution of generative artificial intelligence has significantly increased the realism, accessibility, and scalability of deepfake technologies, posing substantial risks to digital trust, cybersecurity, media authenticity, and societal stability. As synthetic manipulation techniques continue to advance across visual, auditory, and textual domains, traditional unimodal detection approaches have become increasingly insufficient. In response to these emerging challenges, this research presented **DeepGuard**, a comprehensive multimodal deepfake detection framework that integrates image, video, audio, and text analysis through an attention-based fusion mechanism supported by lightweight pretrained architectures.

The proposed framework leverages pretrained MobileNetV2 backbones for spatial feature extraction in image, video frame, and spectrogram-based audio analysis, ensuring efficient yet robust representation learning. For textual analysis, a computationally lightweight Multinomial Naïve Bayes classifier combined with TF–IDF vectorization provides complementary semantic signals. These heterogeneous modality embeddings are projected into a shared latent space and fused using an adaptive attention mechanism that dynamically assigns importance weights based on contextual reliability. This design enables the system to effectively capture cross-modal inconsistencies that often characterize sophisticated deepfake manipulations.

Extensive experimental evaluation demonstrates that multimodal integration significantly enhances detection performance compared to unimodal configurations. The ablation studies validate the importance of each architectural component, particularly the attention-based fusion mechanism, which consistently outperforms static concatenation and fixed-weight strategies. Furthermore, computational complexity analysis confirms that the framework achieves a favorable balance between accuracy and efficiency, largely due to the use of lightweight pretrained backbones and modular design. Deployment analysis further highlights the framework’s suitability for cloud-based, enterprise-level, and edge environments, supporting horizontal scalability, GPU acceleration, and near real-time inference capabilities.

The DeepGuard framework contributes to the field of deepfake detection in several meaningful ways. First, it demonstrates that pretrained lightweight architectures can achieve strong multimodal detection performance without relying on excessively deep or computationally expensive models. Second, it provides an adaptive fusion strategy that improves robustness against missing or noisy modalities. Third, it offers a deployment-ready design that bridges the gap between theoretical research and practical real-world application. By integrating efficiency, scalability, and multimodal reasoning, the proposed approach advances the development of reliable detection systems capable of addressing increasingly complex synthetic media threats.

Nevertheless, the deepfake detection problem remains inherently dynamic, driven by the continuous advancement of generative technologies. As future deepfake methods become more temporally consistent, semantically coherent, and perceptually indistinguishable from authentic content, detection systems must evolve accordingly. Continuous learning mechanisms, enhanced semantic modeling, adversarial robustness strategies, and expanded multimodal datasets will be essential to sustain long-term effectiveness. The DeepGuard framework establishes a strong foundation upon which such future advancements can be built.

In conclusion, this work underscores the importance of multimodal analysis and adaptive fusion in combating synthetic media manipulation. By combining efficient pretrained feature extractors with attention-based integration, DeepGuard delivers a scalable, explainable, and high-performing deepfake detection solution. The framework not only addresses current deepfake challenges but also provides a flexible architectural blueprint for future research aimed at preserving authenticity and trust in the rapidly evolving digital information landscape.

CONFLICT OF INTEREST

The authors declare that they have no known financial interests, personal relationships, institutional affiliations, or competing professional commitments that could have appeared to influence the work reported in this paper. This research was conducted independently, and no external funding or commercial sponsorship was received that could have influenced the study design, data collection, analysis, interpretation of results, or manuscript preparation.

The authors further confirm that there are no conflicts of interest related to intellectual property, employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications, or other financial benefits associated with this research. All results presented in this manuscript are the outcome of unbiased academic investigation conducted in accordance with ethical research standards.

REFERENCES

[1] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[3] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[4] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[5] I. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[6] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018.

[7] H. Nguyen et al., "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," *IEEE International Conference on Biometrics (ICB)*, 2019.

[8] S. Agarwal et al., "Detecting Deep-Fake Videos from Appearance and Behavior," *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

[9] T. Kinnunen et al., "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[10] J. Yamagishi et al., "ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge," *IEEE Journal of Selected Topics in Signal Processing*, 2021.

[11] X. Wang et al., “FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces,” *arXiv preprint arXiv:1909.06122*, 2019.

[12] S. Verdoliva, “Media Forensics and DeepFakes: An Overview,” *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[13] Z. Yan et al., “DeepFakeBench: A Comprehensive Benchmark of DeepFake Detection,” *arXiv preprint arXiv:2307.01426*, 2023.

[14] T. Tan et al., “Learning Audio-Visual Deepfake Detection,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] H. Khalid and S. S. Woo, “OC-FakeDect: Classifying Deepfakes Using One-Class Variational Autoencoder,” *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[16] Y. Zhou and S. Lim, “Joint Audio-Visual Deepfake Detection,” *IEEE International Conference on Computer Vision Workshops*, 2021.

[17] T. Zellers et al., “Defending Against Neural Fake News,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[18] OpenAI, “Detecting AI-Generated Text,” 2019.

[19] S. Uchendu et al., “Authorship Attribution for Neural Text Generation,” *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[20] P. Korshunov and S. Marcel, “Deepfake Detection: Current Challenges and Next Steps,” *IEEE Signal Processing Magazine*, 2022.