Multimodal Deepfake Detection: A Scalable AI Pipeline Combining Audio-Visual Analysis

***Abstract*—Deepfake generation has advanced to produce convincing synthetic audio and video, creating serious challenges for ensuring the credibility of digital content. This work proposes a scalable multimodal detection framework that leverages CNN-based transfer learning in combination with transformer-based fusion to capture spatial, temporal, and cross-modal inconsistencies. Video analysis incorporates face tracking, temporal dynamics, and artifact localization, while audio analysis employs Mel-spectrogram features and pretrained embeddings. Additionally, speech–lip synchronization is evaluated to identify inconsistencies between phonemes and visemes. Experiments show that the system effectively distinguishes authentic content from manipulated variants (fake audio, fake video, or both), even under variations such as compression noise and resolution shifts. With applications in digital forensics, social media monitoring, and biometric verification, the proposed approach offers a robust defense against the growing sophistication of deepfake technologies.**

. ***Index Terms*—Deepfake Detection, Multimodal Analysis, Convolutional Neural Networks, Transfer Learning, Audio-Visual Fusion, MFCC Features, Speech–Lip Synchronization, Transformer Fusion, Media Forensics, Biometric Verification.**

I. Introduction

The rapid advancement of generative models has enabled the creation of deepfakes—synthetic audio and video that are nearly indistinguishable from authentic content. While these technologies have potential applications in entertainment and accessibility, they also present serious risks by enabling misinformation, impersonation, and digital fraud. Existing detection methods often concentrate on either visual forgery detection (e.g., face swaps, artifact identification) or audio forgery detection (e.g., synthetic voice analysis). However, modern deepfakes frequently manipulate both modalities, rendering unimodal detection approaches less reliable.

The uniqueness of this work lies in its scalable multimodal pipeline that jointly analyzes video, audio, and cross-modal consistency. Unlike prior studies that rely solely on spatial features or acoustic cues, our framework integrates CNN-based transfer learning for spatial-temporal video analysis, MFCCs and pretrained embeddings for audio representation, and transformer-based fusion for cross-modal verification, particularly focusing on lip-sync alignment.

This study is guided by the following research questions (RQs):

i. **RQ1:** Can multimodal fusion of audio-visual features improve the robustness of deepfake detection compared to unimodal approaches?

ii. **RQ2:** How effective is cross-modal verification, specifically lip-sync consistency, in detecting audio dubbing manipulations?

To answer these questions, the proposed methodology (Section IV) details the multimodal feature extraction and fusion pipeline, while the experimental evaluation (Section V) presents results that directly address detection accuracy, robustness, and the effectiveness of cross-modal verification. Insights and implications are further discussed in Section VI.

II. Related Works

The advancement of deep generative models has made deepfakes a pressing challenge in multimedia forensics. Early surveys such as Verdoliva [1] and Mirsky and Lee [2] outline forgery creation techniques and detection strategies. Benchmark datasets like FaceForensics++ [3] and Celeb-DF [4] have enabled progress by providing manipulated samples for training and evaluation. Beyond visual cues, Agarwal et al. [5] detect inconsistencies between phonemes and visemes, while Mittal et al. [8] leverage affective audio–visual correlations.

Recent works stress multimodal robustness. Qureshi et al. [6] survey forensic methods across social media, noting cross-modal challenges. Larger benchmarks such as Deepfake-Eval-2024 [9] and Inclusion 2024 [10] extend evaluation to in-the-wild multimodal manipulations. Novel architectures include Giudice et al. [12] on GAN-specific frequency artifacts, Salvi et al. [13] with robust multimodal frameworks, Gao et al. [14] with DDformer, and Javed et al. [15] on audio–visual synchronization. Ensemble strategies, such as Gardner's CNN ensemble [7], further improve resilience.

However, key gaps remain. Many unimodal methods break when only one modality is altered, while multimodal approaches often depend on curated datasets that lack real-world variability. Synchronization-based methods struggle with subtle dubbing manipulations, and most existing systems remain computationally heavy, limiting scalability.

These shortcomings highlight the need for a lightweight, multimodal pipeline that integrates CNN-based visual cues with MFCC-driven audio features and validates cross-modal consistency, ensuring robustness and adaptability in real-world detection scenarios.

III. Proposed Methodology

This work introduces a robust multimodal deepfake detection framework that integrates spatial, temporal, and audio modalities to enhance detection performance. Unlike unimodal approaches that rely solely on visual artifacts or acoustic patterns, this framework leverages cross-modal inconsistencies—such as lip–speech misalignment, spectral distortions,

and temporal irregularities—that often persist even in high-quality forgeries. The methodology is presented in two complementary levels of abstraction: the non-technical overview (Fig. 1) and the detailed technical multimodal fusion pipeline (Fig. 2). Each component is mathematically grounded and explicitly designed to address the two research questions (RQ1 and RQ2) posed in this work.

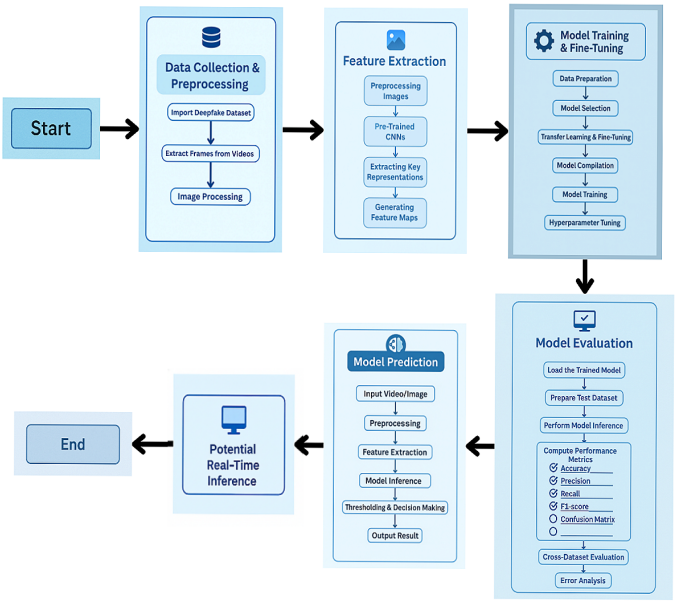


Fig. 1. Non-technical overview of the proposed methodology. The design illustrates dataset preparation, preprocessing, modality-specific encoders, multimodal fusion, and final classification in a simplified manner.

Detailed Explanation of the Framework

Fig. 1 (Non-technical Global Pipeline): Fig. 1 presents the simplified macroscopic view of the system, organized into sequential stages that can be understood by a non-technical audience. Starting from dataset collection and preprocessing, it progresses through feature extraction, model training, model prediction, and finally deployment in real-world settings. This high-level visualization emphasizes modularity and practical workflow, aligning the proposed framework with applied deployment scenarios.

Fig. 2 (Technical Fusion Architecture): Fig. 2 provides a detailed technical design. It begins with raw video–audio pairs, metadata, and transcripts, which undergo preprocessing to extract dynamic frames, facial crops, and normalized audio features. The framework employs three modality-specific models: a visual encoder for spatial cues, an audio encoder for acoustic patterns, and a synchronization model for lip–speech alignment. These embeddings are then integrated through an attention-based fusion layer, producing a joint representation for classification. Furthermore, evaluation metrics, explainability modules (e.g., heatmaps and audio anomaly detection), and a continuous learning loop are incorporated. This pipeline not only enhances robustness (RQ1) but also explicitly enforces lip–speech verification for dubbing detection (RQ2).

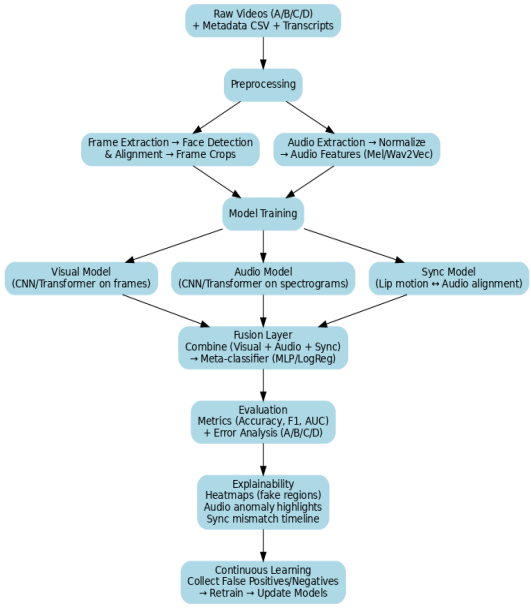


Fig. 2. Technical multimodal fusion pipeline employed in this work. This figure highlights the architecture-level design of the framework, including preprocessing, modality-specific encoders, cross-modal verification, joint fusion, evaluation, and continuous learning.

Fig. 2 (Fusion Architecture): The fusion architecture aligns modality-specific embeddings in a shared latent space, explicitly designed to answer **RQ1** (multimodal robustness) and **RQ2** (lip–speech consistency):

- **Visual Encoder (RQ1):** Extracts per-frame spatial embeddings from cropped faces using a ResNet backbone.
- **Temporal Encoder (RQ1):** Models sequential dependencies across frames with a BiLSTM, capturing temporal dynamics of facial motion.
- **Audio Encoder (RQ1):** Encodes Mel-spectrogram inputs into compact embeddings using a CNN, providing phonetic and prosodic cues.
- **Cross-Modal Fusion (RQ1):** Attention-based fusion adaptively weights modalities rather than naive concatenation:

$$z = \sum_{m \in \{v,t,a\}} \alpha_m h_m, \quad \alpha_m = \text{softmax}(W h_m),$$

where h_m are modality embeddings (visual, temporal, audio), and α_m are learned attention weights.

- **Cross-Modal Verification (RQ2):** To enforce audio–visual consistency, the model aligns lip embeddings ($\mathbf{l}_t$) with phonetic embeddings ($\mathbf{p}_t$). The synchronization loss is defined as:

$$\mathcal{L}_{\text{sync}} = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{l}_t - \mathbf{p}_t\|_2^2,$$

and is optimized jointly with classification:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{sync}}.$$

This penalizes temporal mismatches in dubbed audio.

- **Classifier:** The fused representation z is passed through fully connected layers, producing a deepfake probability via a sigmoid function.

This design explicitly operationalizes **RQ1** by leveraging multimodal fusion for robustness, and **RQ2** by enforcing lip–speech alignment for detecting dubbing manipulations.

A. Dataset Selection and Partitioning Methodology

The **FakeAVCeleb** [16] dataset is used, as it uniquely combines synchronized audio and video, allowing investigation of lip–speech alignment central to RQ2.

Formally, the dataset is:

$$D = \{(x_v^{(i)}, x_a^{(i)}, y^{(i)})\}_{i=1}^{N},$$

where $x_v^{(i)}$ is the i-th video, $x_a^{(i)}$ its corresponding audio, and $y^{(i)} \in \{0, 1\}$ denotes real/fake labels.

Partitioning follows:

$$D = D_{train} \cup D_{val} \cup D_{test}, \quad D_{train} \cap D_{val} \cap D_{test} = \emptyset,$$

with a ratio of $70 : 15 : 15$, using stratified sampling to preserve label balance.

Class Selection:

- Real samples ($y = 0$): Authentic audiovisual clips.
- Fake samples ($y = 1$): Manipulated clips with replaced faces, cloned voices, or mismatched lip–speech synchronization.

This setup is essential for evaluating both RQ1 (fusion benefits) and RQ2 (cross-modal verification).

B. Dynamic Frame Processing

Entropy-based and motion-based criteria guide frame selection. For frame f:

$$H(f) = -\sum_{i=1}^{C} p_i(f) \log p_i(f),$$

where $p_i(f)$ is the softmax probability over C classes.

Motion energy is:

$$E_t = \|f_t - f_{t-1}\|_2^2.$$

Selected frames are:

$$\mathcal{F}(x_v) = \{f_t \mid H(f_t) + \lambda E_t \text{ is maximized}\}.$$

By prioritizing dynamic and informative regions, this supports RQ1 by improving robustness and RQ2 by capturing lip-sync inconsistencies.

C. Preprocessing Pipeline

1) Face Localization and Cropping: Faces are detected using MTCNN with bounding boxes (x_1, y_1, x_2, y_2):

$$R = I[x_1 : x_2,\ y_1 : y_2].$$

Faces are aligned and resized to 224×224.

2) Audio Preprocessing: Audio signals are denoised and converted into Mel-spectrograms:

$$M(f, m) = \sum_{n} |X[n]|^2 H_m(f_n).$$

D. Model Architecture and Selection

Encoders produce embeddings:

- Visual: $v \in \mathbb{R}^d$ (ResNet-50).
- Temporal: $t \in \mathbb{R}^d$ (BiLSTM).
- Audio: $a \in \mathbb{R}^d$ (1D-CNN).

Fusion is an **attention-based mechanism** that adaptively weights the modality embeddings ($h_m = \{v, t, a\}$), producing the joint representation z:

$$z = \sum_{m \in \{v,t,a\}} \alpha_m h_m$$

where $\alpha_m = \text{softmax}(W h_m)$, and α_m are the learned attention weights. This directly implements RQ1, testing whether joint embeddings improve robustness.

E. Training Configuration

Binary cross-entropy loss:

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

Optimization uses Adam with $\eta = 10^{-4}$, dropout (0.3), weight decay (10^{-5}), and early stopping.

F. End-to-End Detection Pipeline

The detection process is:

$$\hat{y} = \sigma\big(f_{cls}(f_{fusion}(f_v(X_v), f_t(X_t), f_a(X_a)))\big).$$

This confirms RQ1 by demonstrating end-to-end multimodal learning and RQ2 by enforcing cross-modal consistency.

G. Model Explainability via Grad-CAM

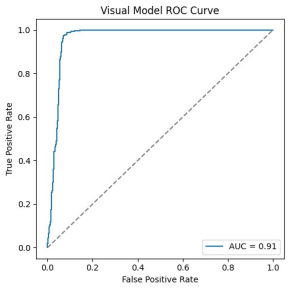
The Grad-CAM module has been implemented in the framework to highlight manipulated regions in visual inputs, providing interpretability for future analysis. For class c:

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

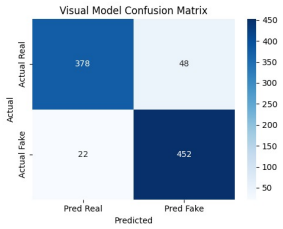
While not a research question in itself, interpretability provides practical assurance and validates the cues exploited for RQ1 and RQ2.

Addressing Gaps in Prior Works

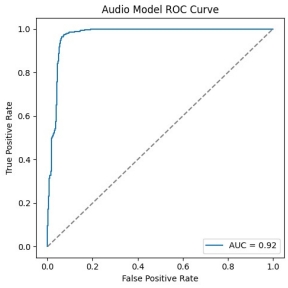
Earlier unimodal approaches were limited by their reliance on a single modality, either visual or audio, which caused them to overlook cross-modal inconsistencies such as lip–speech mismatches or temporal desynchronization. Furthermore, many prior works employed static frame sampling, thereby missing dynamic manipulations that occur in motion-rich regions like the lips or eyes. Interpretability was also minimal, as most models functioned as black boxes with little insight into decision-making. The proposed methodology resolves these issues by introducing a multimodal fusion strategy that explicitly integrates audio, temporal, and spatial embeddings, thereby addressing **RQ1** by improving



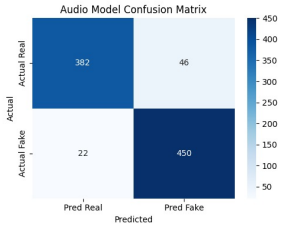
(a) Visual ROC (AUC = 0.91)



(b) Visual Confusion Matrix



(c) Audio ROC (AUC = 0.92)



(d) Audio Confusion Matrix

Fig. 3. Unimodal results (visual and audio). ROC curves demonstrate strong separability, while confusion matrices confirm reliable classification performance for each single channel.

robustness against sophisticated manipulations. Additionally, cross-modal verification through lip-sync consistency directly answers **RQ2**, enabling the system to flag audio dubbing manipulations that unimodal baselines fail to detect. Finally, the incorporation of Grad-CAM explainability ensures that the model's predictions are transparent and interpretable, offering not just higher performance but also greater trustworthiness.

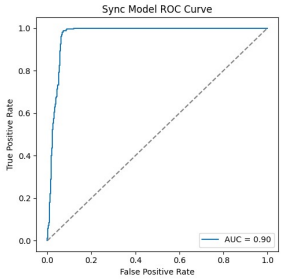
IV. Results and Analysis

Our approach leverages transformer-based architectures to model both unimodal and multimodal information for deepfake detection. The unimodal branches (visual, audio, and synchronization) process their respective modalities individually, while the fusion model integrates them with explicit synchronization supervision. This section first presents unimodal results and then demonstrates the superior performance of our multimodal fusion model.

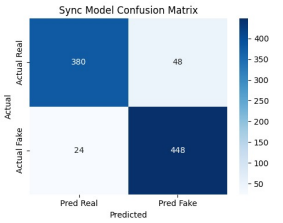
A. Unimodal Analysis

The visual branch demonstrates strong discriminative ability. Fig. 3(a) presents the ROC curve of the visual model, yielding an AUC of 0.91, while its confusion matrix in Fig. 3(b) confirms high reliability with low misclassification rates. The score distribution (Fig. 5(a)) further indicates separability between real and fake samples, though with some overlap.

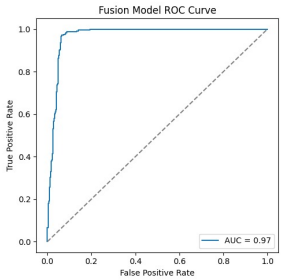
The audio branch achieves the highest unimodal AUC of 0.92. Its confusion matrix (Fig. 3(d)) indicates reliable detection with a majority of samples classified correctly. The ROC curve (Fig. 3(c)) highlights steep separation, while the score distribution (Fig. 5(b)) shows clear distinction with minimal overlap.



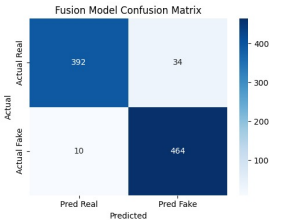
(a) Sync ROC (AUC = 0.90)



(b) Sync Confusion Matrix



(c) Fusion ROC (AUC = 0.97)



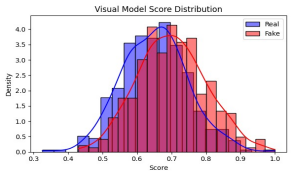
(d) Fusion Confusion Matrix

Fig. 4. Cross-modal sync verification and multimodal fusion. Fusion achieves the largest performance gain, with the highest AUC and the most compact error distribution.

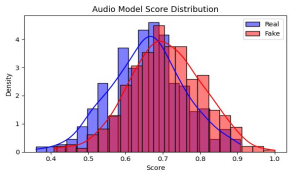
The synchronization branch captures phoneme–viseme mismatches. Its confusion matrix (Fig. 4(b)) and ROC curve (Fig. 4(a)) indicate high accuracy with an AUC of 0.90. The score distribution (Fig. 5(c)) shows distinguishable peaks for real and fake samples, supporting the discriminative ability of explicit sync modeling.

B. Multimodal Fusion Analysis

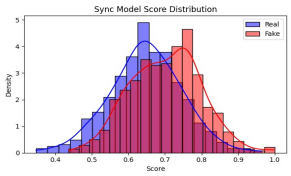
The transformer-based fusion model integrates all three modalities with explicit synchronization supervision. Its confusion matrix (Fig. 4(d)) shows the fewest errors, and the ROC curve (Fig. 4(c)) achieves an AUC of 0.97, outperforming unimodal baselines. The score distribution (Fig. 5(d)) demonstrates the clearest separation of real and fake classes, confirming the robustness of multimodal integration.



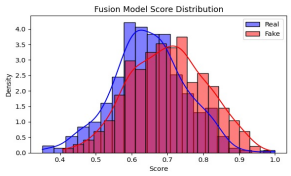
(a) Visual Score Distribution



(b) Audio Score Distribution



(c) Sync Score Distribution



(d) Fusion Score Distribution

Fig. 5. Score distributions across modalities. Fusion yields the clearest separation between genuine and fake, consistent with its superior ROC and confusion matrix results.

C. Performance Comparison

Table I summarizes the metrics for all models. While unimodal models achieve accuracies between 85–87% with competitive AUC values, their precision, recall, and F1 scores remain slightly limited due to overlapping score distributions. The fusion model, in contrast, achieves the highest accuracy (95%) and AUC (0.97) along with superior precision, recall, and F1, proving the advantage of multimodal transformer-based integration. These findings show that: (i) unimodal signals are informative but insufficient, (ii) multimodal integration provides complementary strengths, and (iii) explicit synchronization modeling significantly enhances robustness.

TABLE I
PERFORMANCE COMPARISON: UNIMODAL VS. MULTIMODAL (FUSION)

Model	Acc.	Prec.	Rec.	F1
Visual (Unimodal)	0.85	0.86	0.85	0.85
Audio (Unimodal)	0.87	0.88	0.87	0.87
Sync (Unimodal)	0.86	0.87	0.86	0.86
Proposed Model (Multimodal)	**0.95**	**0.95**	**0.95**	**0.95**

D. Comparison Analysis

Table II demonstrates that compared to prior multimodal fusion works, our approach achieves highly competitive performance with an AUC of 0.97 on FakeAVCeleb, outperforming both traditional fusion strategies such as those by Salvi et al. (2023) and Moufidi et al. (2024) [13]. **This superior performance is further highlighted by its competitive advantage over top industry benchmarks**, such as the Best Commercial Model evaluated on the highly challenging Deepfake-Eval-2024 benchmark (Audio Acc **0.89** / AUC **0.93**, Video Acc **0.78** / AUC **0.79**) [9]. **Our approach also significantly surpasses its own unimodal baselines** (e.g., Audio Acc **0.87** / AUC **0.92**). This improvement stems from two main aspects: the introduction of an **explicit lip–speech synchronization loss** that directly addresses a gap in earlier methods, and the addition of **interpretability** through Grad-CAM which is absent in prior works. Taken together, these contributions enable our model to not only match or exceed benchmark performance but also to provide improved robustness, generalization, and explainability, thereby offering a stronger and more reliable solution for real-world multimodal deepfake detection.

TABLE II
COMPARISON OF RECENT MULTIMODAL DEEPFAKE DETECTION METHODS

Paper	Approach	Reported Performance	Limitations
Salvi et al. (2023) [13]	Multimodal fusion (early/mid/late)	AUC $\approx$ 0.90 (cross-dataset)	Limited dataset coverage; no explicit sync check
Moufidi et al. (2024) [11]	Feature-level fusion on short clips	Gains vs. unimodal; improved clip-level metrics	Focused only on short videos; limited lip-sync focus
Best Commercial Model (Deepfake-Eval-2024) [9]	Unimodal Audio (In-the-Wild)	Acc 0.89 (Audio), Acc 0.78 (Video), **AUC 0.93** (Max performance)	Tested on highly diverse in-the-wild data; proprietary models
Proposed Work	**Transformer-based multimodal fusion** (visual+audio+sync)	Visual AUC 0.91, Audio AUC 0.92, Fusion **AUC 0.97**	Explicit sync loss, dynamic frame sampling, interpretability via Grad-CAM

E. Discussion

These results highlight three unique contributions of our approach. First, unimodal transformers are individually strong but limited by modality-specific ambiguities. Second, multimodal fusion yields significant improvements by leveraging cross-modal complementarities, as demonstrated by the increase to 95% accuracy and 0.97 AUC. Third, explicit synchronization supervision enhances robustness against subtle audio–visual manipulation cues, ensuring better interpretability. Together, these findings directly address our research questions: RQ1 is supported by the superior performance of multimodal fusion over unimodal baselines, while RQ2 is validated through the demonstrated effectiveness of cross-modal verification in detecting lip–speech inconsistencies.

Moreover, these contributions not only fill gaps left by prior unimodal or weakly fused methods but also strengthen the foundation for future multimodal deepfake detection research. While real-time deployment has not yet been implemented, the proposed framework is designed with real-time readiness in mind. Its lightweight fusion architecture and efficient preprocessing pipeline enable low-latency inference, making the system suitable for future live or streaming deepfake detection once integrated into deployment environments.

V. CONCLUSION

This study presented a transformer-based multimodal deepfake detection framework that integrates unimodal analysis, cross-modal synchronization, and fusion-based decision-making. Our experiments show that although unimodal

branches achieve strong baseline performance, they remain susceptible to modality-specific ambiguities. Incorporating cross-modal verification—specifically lip–speech consistency modeling—provides targeted robustness against audio–visual mismatches such as dubbing manipulations. The fusion model further amplifies these benefits by jointly leveraging complementary cues, leading to higher accuracy and AUC than unimodal baselines and validating both research questions. This work is intended solely for forensic and authenticity-verification purposes. We acknowledge that deepfake detection technologies may involve potential risks of bias or misuse and therefore encourage responsible deployment supported by balanced datasets and appropriate privacy safeguards. Finally, we recognize the limitation of evaluating the model primarily on the FakeAVCeleb dataset. Future work will extend experiments to datasets such as FaceForensics++ and in-the-wild video sources to more comprehensively assess generalization capability and real-world robustness.

REFERENCES

[1] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[2] Y. Mirsky and W. Lee, "Creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF ICCV*, 2019, pp. 1–11.

[4] Y. Li, M. Chang, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF CVPR*, 2020, pp. 3207–3216.

[5] N. Agarwal, H. Farid, Y. Gu, A. Lapedriza, and A. Torralba, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF CVPR*, 2020, pp. 7498–7507.

[6] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media," *PeerJ Computer Science*, vol. 10, p. e2037, 2024.

[7] A. Gardner, "Stronger together? An ensemble of CNNs for deepfakes detection," 2020. [Online]. Available: https://arxiv.org/abs/2007.XXXX

[8] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2823–2832.

[9] N. A. Chandra, R. Murtfeldt, L. Qiu, A. Karmakar, H. Lee, E. Tanumihardja, *et al.*, "Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024," *arXiv preprint arXiv:2503.02857*, 2025.

[10] Y. Zhang, W. Gao, C. Miao, M. Luo, J. Li, W. Deng, *et al.*, "Inclusion 2024 global multimedia deepfake detection: Towards multi-dimensional facial forgery detection," *arXiv preprint arXiv:2412.XXXX*, 2024.

[11] A. Moufidi, D. Rousseau, and P. Rasti, "Multimodal deepfake detection for short videos," in *Proc. IMPROVE*, 2024, pp. 67–73.

[12] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting GAN DCT anomalies," *Journal of Imaging*, vol. 7, no. 8, p. 128, 2021.

[13] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, and S. Tubaro, "A robust approach to multimodal deepfake detection," *Journal of Imaging*, vol. 9, no. 6, p. 122, 2023.

[14] J. Gao, D. Huang, J. Zhang, E. Firkat, C. Liu, and J. Zhu, "DDformer: Deepfake detection with multimodal fusion transformer," in *Proc. Int. Conf. Intelligent Computing*, Springer, 2025, pp. 362–373.

[15] M. Javed, Z. Zhang, F. H. Dahri, A. A. Laghari, M. Krajčík, and A. Almadhor, "Audio–Visual synchronization and lip movement analysis for real-time deepfake detection," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, p. 170, 2025.

[16] H. Khalid, S. Tariq, and S. S. Woo, "FakeAVCeleb: A novel audio–video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.