

AV-DeepFake-Net: Attention-Guided and Uncertainty-Aware Network for Audiovisual DeepFake Detection

Nasir Saleem¹ , Ahmad Ali^{1*} , Zhuoqi Zeng¹ , Adeel Hussain² , Sami Bourouis³ , Sami Dhahbi⁴ , Afef Dhahbi⁵ 

¹ Department of Computer Science and Technology, Hainan Bielefeld University of Applied Sciences, Hainan (China)

² School of Computing, Engineering and the Built Environment, Edinburgh Napier University, 10 Colinton Rd, Edinburgh EH10 5DT (UK)

³ Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944 (Saudi Arabia)

⁴ Applied College of Mahail Aseer, King Khalid University, Muhayil Aseer 62529 (Saudi Arabia)

⁵ Department of Financial and Accounting Management Programs, Applied College of Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, (Saudi Arabia)

* Corresponding author: ali.ahmad@hainan-biuh.edu.cn

Received 30 January 2026 | Accepted 21 July 2026 | Published 28 August 2026



ABSTRACT

Detecting audio-visual DeepFakes (AV-DeepFakes) has become increasingly critical with the rapid proliferation of accessible synthetic media generation tools across consumer platforms. In this work, we propose a high-performance, deployment-efficient AV-DeepFake detection framework tailored for real-world consumer devices. The proposed model integrates a 3D convolutional visual encoder with a 2D convolutional audio encoder to learn synchronized multimodal representations, effectively capturing spatial, spectral, and prosodic inconsistencies inherent in manipulated content. To detect temporal forgeries, we introduce a bidirectional complementary boundary module that precisely localizes manipulation onsets and offsets. A cross-modal attention fusion mechanism aggregates modality-specific cues, while an uncertainty-aware gating strategy suppresses unreliable signals to improve robustness. Furthermore, a cross-modal discrepancy minimization loss encourages alignment for genuine samples while maximizing divergence for forged content, strengthening multimodal consistency learning. Extensive evaluations on FaceForensics++ and LAV-DF demonstrate the effectiveness of the proposed approach, achieving 97.1% AUC for clip-level detection and an 81.2% F1-score for temporal boundary localization, while reducing inference time by 5× compared with transformer-based methods.

KEYWORDS

AV DeepFake Detection, Cross-Modal Fusion, Temporal Boundary Localization, Uncertainty-Aware Modelling.

DOI: 10.9781/ijimai.2026.6431
eLocator: e6431

I. INTRODUCTION

AV-DEEPFAKE is an increasingly sophisticated form of synthetic media enabled by advances in generative models, where speech and facial expressions are manipulated to realistically imitate individuals [1]. While early detection focused on visual artifacts, modern multimodal DeepFakes tightly synchronize audio and lip movements, making verification significantly more challenging and undermining public trust [2]. The widespread availability of consumer-grade tools and open-source pipelines has further lowered the barrier to creating high-quality forgeries [3], highlighting the need

for detection frameworks that are accurate, efficient, and suitable for deployment on resource-constrained devices [4].

Existing studies [5]–[7] largely rely on visual cues, using CNN-based or hybrid architectures to detect GAN-generated content [8]–[11]. Although effective for image-level analysis, these approaches often fail to capture temporal and cross-modal inconsistencies [12]. Multimodal methods address this by integrating audio-visual information [13], [14], but simple fusion strategies are often suboptimal [15]. Recent approaches improve robustness through cross-modal learning and generalization strategies [16]–[18], yet many overlook temporal boundary localization or depend on computationally expensive transformer-based models.

Benchmark datasets such as FaceForensics++ [19], Celeb-DF [20], and DFDC [21] provide valuable evaluation resources, while multimodal datasets like FakeAVCeleb [22] and LAV-DF [23] enable joint audio-visual and temporal analysis. Despite these advances, existing methods [24]-[26] remain limited by high computational cost and sensitivity to real-world noise [27]. Although frameworks such as BA-TFD [23] highlight the importance of temporal boundary modelling, they are not optimized for lightweight deployment.

To address these gaps, we propose a compact and efficient AV-DeepFake detection framework that jointly models temporal consistency, cross-modal alignment, and uncertainty-aware fusion. Built on lightweight 3D visual and 2D audio encoders, the model captures spatiotemporal and spectral cues while enabling real-time inference. A bidirectional boundary module improves temporal localization, and an uncertainty-aware fusion gate adaptively weights modality reliability. Additionally, a cross-modal discrepancy minimization (CMDM) loss enhances alignment for genuine samples and separation for manipulated ones, leading to robust and calibrated multimodal decisions.

1. We propose a computationally efficient audio-visual DeepFake detection framework that jointly models spatial, spectral, and temporal inconsistencies using compact 3D and 2D convolutional encoders. The enhanced boundary module, equipped with a bidirectional complementary boundary generator, enables precise localization of forgery onsets and offsets while maintaining low latency suitable for real-time inference on consumer devices.
2. We design a dual fusion mechanism that combines attention-based cross-modal correlation with entropy-guided uncertainty weighting, allowing the system to adaptively suppress unreliable modalities under noisy or compressed conditions. In addition, a cross-modal discrepancy minimization (CMDM) loss, implemented via modality-specific and fused clip-level heads, enforces prediction agreement for genuine clips and controlled divergence for manipulated ones, improving calibration and interpretability across audio-visual modalities.
3. Extensive experiments on the LAV-DF and FaceForensics++ datasets demonstrate that the proposed framework achieves superior performance compared to unimodal and existing multimodal baselines, with AUC = 97.1% for clip-level detection and temporal Boundary F1 = 81.2% for manipulation localization. Moreover, our model reduces inference latency by up to $5\times$ compared to transformer-based baselines and maintains strong performance under realistic degradations such as motion blur, aggressive video compression, and noisy audio, supporting its suitability for on-device media authentication.

The remainder of the paper is organized as follows. Section II details the proposed architecture. Section III describes the experimental setup, datasets, and evaluation metrics. Section IV provides in-depth performance analysis and ablation studies. Section V discusses the main findings. Section VI provides the limitations of the model. Finally, Section VII concludes the study.

II. METHODOLOGY: AV DEEPFAKE MODEL

The proposed audio-visual DeepFake detection framework builds on [23], introducing key enhancements for improved robustness and real-world deployment. It employs lightweight encoders for efficient, real-time operation on resource-constrained devices, and an uncertainty-aware fusion mechanism that adaptively weights audio-visual cues based on entropy. A bidirectional temporal consistency constraint strengthens boundary localization, while a clip-level detection head with cross-modal consistency and disagreement losses jointly optimizes

localization and classification. These improvements transform into a practical and deployable multimodal detection system [23].

Fig. 1 illustrates the architecture of the proposed AV-DeepFake-Net framework. The model consists of two parallel modality-specific streams and several shared processing modules. The visual stream processes input video frames using a 3D convolutional encoder to extract temporally structured spatiotemporal features. In parallel, the audio stream processes synchronized audio signals converted into log-mel spectrograms using a 2D convolutional encoder, capturing spectral and prosodic patterns. The extracted features are then passed to the temporal boundary modeling module, which estimates frame-level authenticity and generates temporal boundary maps indicating potential manipulation regions. This module includes the complementary boundary generator, which refines start and end predictions to improve temporal localization. Next, the cross-modal fusion module integrates audio and visual information. This stage first applies attention-based fusion to capture correlations between modalities, followed by uncertainty-aware weighting that dynamically adjusts the contribution of each modality based on prediction reliability and averaged them. Finally, the classification head aggregates the fused temporal evidence using a learnable pooling mechanism to produce a clip-level DeepFake prediction. The overall design enables the model to jointly reason about spatial, temporal, and cross-modal inconsistencies.

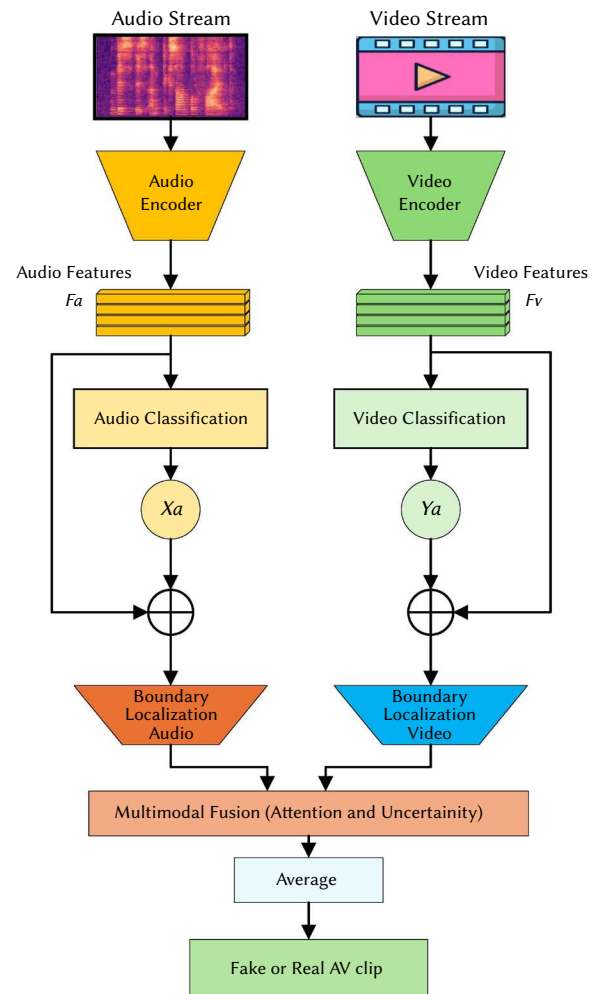


Fig. 1. The architecture of the AV-DeepFake detection framework integrates spatiotemporal encoders, temporal boundary modeling, and uncertainty-aware fusion.

A. Audio-Visual Feature Extraction

1. Visual Encoder

The video stream is processed using a lightweight 3D convolutional backbone. Given an input sequence of RGB video frames denoted as

$$V \in \mathbb{R}^{B \times 3 \times T \times H \times W}, \quad (1)$$

where B denotes the batch size, T represents the temporal length, and (H, W) denote the spatial dimensions. The encoder extracts temporal features,

$$F_v \in \mathbb{R}^{B \times C_f \times T}, \quad (2)$$

where C_f denotes the feature dimensionality.

The video encoder employs a 3D convolutional architecture to jointly capture spatial and temporal patterns from raw video input. Spatial resolution is progressively reduced while preserving temporal dynamics, enabling the model to capture motion cues and fine temporal inconsistencies, such as lip-jaw asynchrony, frame blending, and local facial warping, which are indicative of DeepFake artifacts. The resulting embeddings F_v retain sufficient temporal fidelity for subsequent classification and boundary modeling while remaining computationally efficient for real-time deployment.

2. Audio Encoder

The temporally synchronised audio signal is first transformed into a log-mel spectrogram:

$$A' \in \mathbb{R}^{B \times 1 \times F_m \times T_a}, \quad (3)$$

where F_m denotes the number of mel-frequency bins and T_a represents the temporal length of the audio segment. A lightweight 2D convolutional encoder is then used to extract temporal audio embeddings:

$$F_a \in \mathbb{R}^{B \times C_f \times T}, \quad (4)$$

where temporal alignment with the visual features is achieved through appropriate resampling or interpolation. The audio encoder captures key spectral and temporal characteristics such as formants, harmonics, and prosodic variations. The embeddings provide complementary cues to the visual stream, strengthening multimodal reasoning and improving deepfake detection performance.

B. Frame-Level Authenticity Estimation

To capture fine-grained temporal information, the encoded features from each modality are passed through a frame-wise prediction layer. This layer estimates the probability of each frame as authentic or manipulated. The outputs for video and audio modalities are computed as

$$Y_v = \sigma(\text{Conv}_v(F_v)), \quad Y_a = \sigma(\text{Conv}_a(F_a)), \quad (5)$$

where $\text{Conv}_v(\cdot)$ and $\text{Conv}_a(\cdot)$ denote (1×1) temporal convolution layers, and $\sigma(\cdot)$ is the sigmoid activation that produces frame-level authenticity probabilities:

$$Y_v, Y_a \in \mathbb{R}^{B \times 1 \times T}. \quad (6)$$

These per-frame predictions serve as early indicators of potential manipulation and are concatenated with the corresponding feature sequences to enhance temporal context awareness. The resulting boundary-aware feature representations are expressed as

$$X_v = \text{Concat}(F_v, Y_v), \quad X_a = \text{Concat}(F_a, Y_a), \quad (7)$$

The channel-wise concatenation $\text{Concat}(\cdot, \cdot)$ results in

$$X_v, X_a \in \mathbb{R}^{B \times (C_f+1) \times T}. \quad (8)$$

This operation enriches the feature representation by explicitly incorporating preliminary authenticity cues, allowing subsequent boundary modeling modules to focus on temporally salient regions exhibiting abnormal dynamics or cross-modal inconsistencies.

C. Temporal Boundary Modeling

To localize manipulated regions and capture temporal inconsistencies between modalities, the proposed framework employs specialised boundary modules that extend the temporal proposal matching paradigm inspired by BSN++ [23]. Each module operates on boundary-aware feature sequences X_v and X_a to generate boundary maps:

$$(M_v, M_a) \in \mathbb{R}^{B \times D \times T}, \quad (9)$$

where D denotes the maximum proposal duration and represents the range of possible manipulation lengths. The boundary estimation process follows a multi-stage design for efficient spatiotemporal reasoning on edge devices. First, a boundary matching layer generates proposal tensors through uniform temporal sampling over segment durations $d \in [1, D]$. This results in a feature volume:

$$\mathbb{R}^{B \times n_f \times N \times D \times T}, \quad (10)$$

where n_f denotes the feature dimension and N is the number of sampling points per duration. This representation encodes temporal patterns corresponding to candidate manipulation intervals. A series of three-dimensional and two-dimensional convolutional operations is then applied to compress this proposal volume into boundary likelihood maps:

$$M_v = \phi_v(X_v), \quad M_a = \phi_a(X_a), \quad (11)$$

where $\phi_v(\cdot)$ and $\phi_a(\cdot)$ denote modality-specific boundary extraction functions composed of convolutional layers, nonlinear activations, and sigmoid normalization. To enhance temporal precision, a complementary boundary generator based on a Nested U-Net architecture is employed to refine the boundary activations. This module utilises a hierarchical encoder-decoder structure with skip connections for multi-scale feature aggregation. It produces start and end probability maps:

$$S, E \in \mathbb{R}^{B \times T}, \quad (12)$$

representing the likelihood of manipulation onset and offset, respectively. To improve temporal consistency, a bidirectional processing strategy is introduced. The input sequences are processed in both forward and reverse temporal order, producing (S_f, E_f) and (S_b, E_b) . These are fused using geometric mean normalization,

$$\tilde{S} = \sqrt{S_f \odot S_b}, \quad \tilde{E} = \sqrt{E_f \odot E_b}, \quad (13)$$

where $\odot$ denotes element-wise multiplication. This bidirectional formulation enforces temporal coherence and improves robustness under desynchronisation or abrupt motion variations. To account for modality reliability, an uncertainty-aware mechanism is incorporated. The entropy of each boundary map is used to estimate its confidence:

$$\alpha_v = \exp(-\lambda H(M_v)), \quad \alpha_a = \exp(-\lambda H(M_a)), \quad (14)$$

where $H(\cdot)$ denotes the normalized entropy and λ controls sensitivity to uncertainty. Higher entropy indicates less reliable predictions, resulting in reduced contribution during fusion. This mechanism enables the model to downweight degraded modalities (e.g., noisy audio or blurred video) while emphasizing more reliable temporal cues. To ensure efficiency on consumer devices, the boundary module incorporates channel reduction and grouped convolution operations, significantly reducing parameter count and computational cost without sacrificing performance. The resulting representations

M_v , M_a , and refined boundary maps ($\tilde{S}$, $\tilde{E}$) provide rich, temporally localized evidence of potential manipulations. These are subsequently passed to the cross-modal fusion module for uncertainty-aware integration and final DeepFake classification.

D. Cross-Modal Fusion

To integrate complementary temporal evidence from both modalities, the proposed framework employs a dual-stage fusion strategy consisting of an attention-based fusion mechanism and an uncertainty-aware refinement. This design enables adaptive information weighting between the audio and visual streams, ensuring robustness under varying signal quality conditions frequently encountered in consumer-grade recordings. The baseline fusion mechanism performs cross-modal attention over boundary maps and intermediate feature representations to learn adaptive fusion weights. Given modality-specific boundary maps $M_v, M_a \in \mathbb{R}^{B \times D \times T}$ and encoded features $F_v, F_a \in \mathbb{R}^{B \times C_f \times T}$, the fusion weights w_v and w_a are obtained via an attention function $\Psi(\cdot)$ that captures both intra- and inter-modal dependencies:

$$w_v, w_a = \Psi(F_v, F_a, M_v, M_a), \quad (15)$$

where $\Psi(\cdot)$ consists of a series of convolutional and soft-attention layers that compute correlation scores between modalities based on spatial-temporal context and boundary activation similarity. These attention coefficients are normalized through a softmax function along the modality dimension, producing weights that dynamically reflect the reliability of each modality's evidence at every temporal location.

The fused boundary representation M_f is then computed as a weighted aggregation of modality-specific maps:

$$M_f = w_v \odot M_v + w_a \odot M_a, \quad (16)$$

where $\odot$ denotes element-wise multiplication. This formulation allows the network to learn modality importance directly from data, enhancing the sensitivity to coherent multimodal cues (e.g., synchronized lip motion and speech) while suppressing inconsistent or spurious modality signals. While attention-based fusion effectively captures modality correlations under ideal conditions, real-world signals from consumer devices often suffer from compression artifacts, background noise, or motion blur, which can introduce unreliability in one or both modalities. To mitigate this, an additional uncertainty-aware fusion mechanism is used to modulate each modality's contribution based on its prediction entropy.

For each modality, the entropy of its boundary map distribution is computed over the duration dimension as

$$H(M_i) = -\frac{1}{DT} \sum_{t=1}^T \sum_{d=1}^D \widehat{M}_{i,d,t} \log(\widehat{M}_{i,d,t} + \epsilon). \quad (17)$$

where $i \in \{v, a\}$, $M_{i,d}$ represents the normalized activation for duration d , D is the total number of durations, and ϵ is a small constant for numerical stability. The resulting entropy values quantify uncertainty in boundary predictions—higher entropy indicates lower confidence. Entropy-derived reliability weights α_v and α_a are then computed as

$$\alpha_v = \exp(-\lambda H(M_v)), \quad \alpha_a = \exp(-\lambda H(M_a)), \quad (18)$$

where λ is a scaling factor controlling sensitivity to uncertainty. These weights act as confidence gates, attenuating unreliable modality responses (e.g., degraded audio segments or blurred facial regions) and amplifying more stable temporal evidence. The final uncertainty-weighted fusion output is expressed as

$$\widehat{M}_f = \frac{\alpha_v M_v + \alpha_a M_a}{\alpha_v + \alpha_a}. \quad (19)$$

This adaptive formulation improves performance under real-world conditions, ensuring that the fused representation $\widehat{M}_f$ remains stable even when one modality degrades significantly.

E. DeepFake Classification

Following boundary estimation and multimodal fusion, the model aggregates frame-level authenticity evidence into a clip-level decision using a learnable temporal pooling and gating mechanism. The fused boundary representation $\widehat{M}_f \in \mathbb{R}^{B \times D \times T}$ encapsulates localized evidence of potential manipulation across temporal positions and durations. Instead of performing naive averaging, the model uses AVClipHead, which employs a learnable soft top- k pooling operator that selectively emphasizes high-confidence frames while maintaining differentiability. Given a flattened temporal representation $\mathbf{m}_f \in \mathbb{R}^T$, the pooled score s is computed as

$$s = \text{Softmax}\left(\frac{\mathbf{m}_f}{\tau}\right)^\top \mathbf{m}_f, \quad (20)$$

where τ is a learnable temperature parameter controlling the sharpness of attention over time. Smaller values of τ cause the pooling operation to focus on top-ranked frames (hard top- k behavior), whereas larger values result in smoother averaging. This formulation enables the network to adaptively aggregate evidence from the most informative temporal regions corresponding to manipulation intervals while suppressing neutral segments.

To improve temporal coherence, an optional gating mechanism derived from the complementary boundary generator outputs is applied. Let $\tilde{S}, \tilde{E} \in \mathbb{R}^{B \times T}$ denote the bidirectionally refined start and end probability maps from Section II.C. A temporal gate G is constructed as

$$G = \frac{1}{2}(\tilde{S} + \tilde{E}), \quad (21)$$

which highlights temporally consistent regions where both onset and offset boundaries exhibit high agreement. The gated boundary map $\overline{M}_f$ is obtained through element-wise multiplication:

$$\overline{M}_f = G \odot \widehat{M}_f. \quad (22)$$

To ensure coherent multimodal decision-making, the framework includes modality-specific classification heads for audio (a_head) and visual (v_head) streams. Each head produces a clip-level logit s_a and s_v using soft top- k pooling applied to the corresponding boundary maps M_a and M_v . In parallel, the fused boundary representation $\overline{M}_f$ yields the joint prediction s_{av} through the main AVClipHead. To regularize the relationship between these outputs, a cross-modal discrepancy minimization strategy is employed [28]. This mechanism enforces consistency for genuine (real) clips and controlled divergence for manipulated (fake) clips, encouraging the network to rely on cross-modal coherence during inference. The objectives of this strategy have two complementary terms:

1. **Consistency Loss:** For authentic clips, audio and visual predictions are encouraged to align closely with the joint fused prediction.

$$\mathcal{L}_{\text{cons}} = \|s_{av} - s_a\|_2^2 + \|s_{av} - s_v\|_2^2. \quad (23)$$

2. **Disagreement Loss:** For fake clips, a margin-based loss promotes sufficient separation between the modality-specific predictions to account for cross-modal inconsistencies typically present in DeepFakes:

$$\mathcal{L}_{\text{disc}} = \max(0, m - |s_a - s_v|), \quad (24)$$

where m denotes a predefined margin threshold. The overall clip-level classification loss combines these terms with a binary cross-entropy loss for authenticity prediction:

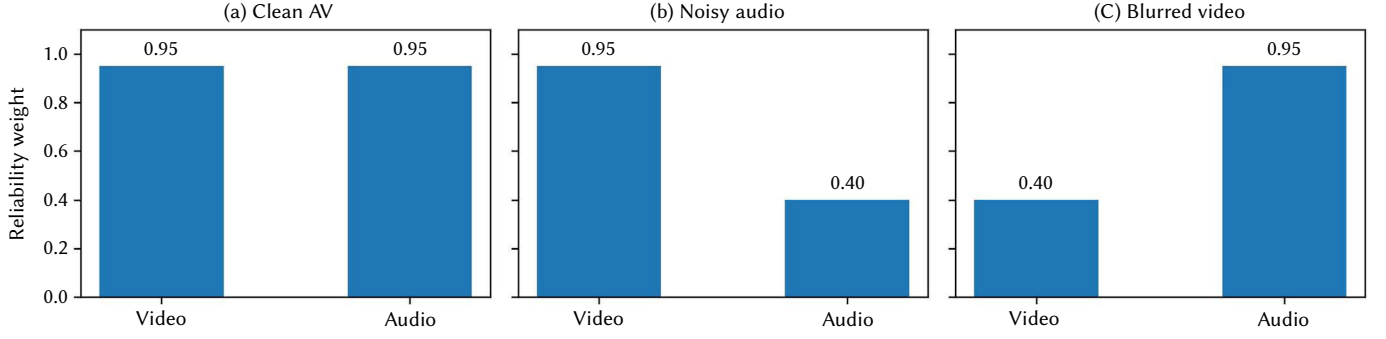


Fig. 2. Conceptual illustration of entropy-based uncertainty weighting. When a modality is degraded (e.g., blur or noise), its boundary entropy increases, leading to lower reliability weight α_i and reduced influence in the fused map $\widehat{M}_f$.

TABLE I. PSEUDOCODE-STYLE SUMMARY OF THE PROPOSED AV-DEEPFAKE-NET FRAMEWORK

Input: Video clip V , synchronized audio signal A

Output: Clip-level prediction s_{av} , fused boundary map $\widehat{M}_f$

1. Extract face-centred RGB frames from V and resize to fixed spatial dimensions.
2. Convert audio signal A into log-mel spectrogram representation A' .
3. Encode visual stream using lightweight 3D CNN to obtain temporal visual embeddings F_v .
4. Encode audio stream using lightweight 2D CNN to obtain temporal audio embeddings F_a .
5. Apply frame-wise logistic heads to compute authenticity probabilities Y_v and Y_a .
6. Concatenate features with frame-level logits to form boundary-aware streams $X_v = [F_v; Y_v]$ and $X_a = [F_a; Y_a]$.
7. Pass X_v and X_a through modality-specific boundary modules to generate boundary maps M_v and M_a .
8. Refine temporal start and end predictions using the complementary boundary generator in both forward and backward directions.
9. Fuse forward and backward start/end maps using geometric mean to obtain $\widetilde{S}$ and $\widetilde{E}$.
10. Compute entropy of M_v and M_a and derive uncertainty-aware reliability weights α_v and α_a .
11. Apply cross-modal attention to estimate modality fusion weights w_v and w_a .
12. Generate fused temporal representation by combining modality-specific boundary maps.
13. Refine fused boundary representation using uncertainty-aware weighting to obtain $\widehat{M}_f$.
14. Construct temporal gate G from refined boundary maps $\widetilde{S}$ and $\widetilde{E}$.
15. Apply gate G to $\widehat{M}_f$ to emphasize temporally consistent manipulation regions.
16. Aggregate gated evidence through AVClipHead using soft top- k pooling to obtain clip-level score s_{av} .
17. Compute modality-specific clip scores s_a and s_v from audio and visual branches.
18. Optimize the Network using binary cross-entropy loss together with CMDM consistency/disagreement regularization.

$$\mathcal{L}_{\text{clip}} = \mathcal{L}_{\text{BCE}}(s_{av}, y) + \lambda_c \mathcal{K}_{\{y=1\}} \mathcal{L}_{\text{cons}} + \lambda_d \mathcal{K}_{\{y=0\}} \mathcal{L}_{\text{disc}}. \quad (25)$$

where $y \in \{0, 1\}$ is the ground-truth authenticity label and λ_c, λ_d are weighting coefficients. This formulation ensures that the model jointly optimizes both detection accuracy and multimodal reliability. The CMDM loss is designed to explicitly model the relationship between audio and visual modalities under both genuine and manipulated conditions. For authentic clips, audio and visual streams originate from the same source and are naturally consistent. The consistency loss therefore encourages the modality-specific predictions to align closely with the fused prediction, reinforcing temporal coherence across modalities. In contrast, DeepFake manipulations often introduce subtle inconsistencies between audio and visual signals, such as lip-sync errors, temporal misalignment, or unnatural prosody. The disagreement loss exploits this property by encouraging a controlled divergence between modality-specific predictions in fake clips. As a result, the model learns to amplify cross-modal inconsistencies in manipulated regions while preserving agreement in genuine segments. This dual objective improves both detection accuracy and temporal consistency, as the model becomes sensitive to modality agreement patterns that are inherently linked to authentic or manipulated content.

F. Uncertainty-Aware Fusion Weights

Fig. 2 provides a conceptual illustration of how the uncertainty-aware fusion mechanism dynamically adjusts the modality reliability weights α_v (video) and α_a (audio) under different signal conditions. In the clean setting (Panel a), both audio and video streams produce low-entropy boundary maps, resulting in high confidence and approximately equal reliability weights ($\alpha_v \approx \alpha_a \approx 1$). When audio is corrupted by noise (Panel b), the boundary estimates for the audio encoder become more uncertain, increasing entropy and consequently reducing its reliability weight ($\alpha_a \downarrow$). The fusion module therefore gives greater emphasis to the video modality ($\alpha_v \uparrow$) when forming the fused boundary representation $\widehat{M}_f$. Conversely, when the video stream is degraded by blur or compression (Panel c), the video boundary confidence decreases, causing the reliability weight for video to drop ($\alpha_v \downarrow$) while the audio weight increases ($\alpha_a \uparrow$). Table I summarizes the inference and training of the proposed AV-DeepFake-Net framework for clarity and reproducibility.

TABLE II. COMPONENT-WISE ABLATION ANALYSIS ON FACEFORENSICS++ AND LAV-DF. CBG: COMPLEMENTARY BOUNDARY GENERATOR, CMDM: CROSS-MODAL DISCREPANCY MINIMIZATION

Configuration	FF++ AUC (%)	FF++ F1 (%)	LAV-DF mAP@0.5 (%)
Baseline (visual only)	91.8	72.5	65.3
+ Audio integration	93.8	74.1	68.1
+ Attention-based fusion	94.6	75.9	70.9
+ Bidirectional CBG	96.5	79.6	74.1
+ Uncertainty-aware fusion	96.7	80.1	76.4
+ CMDM loss (Full model)	97.1	81.2	78.5

III. EXPERIMENTAL SETUP

A. Datasets

We evaluate the proposed framework on two public audio-visual DeepFake datasets that provide complementary supervision for temporal localization and clip-level classification. For temporal boundary localization, we use the LAV-DF dataset¹ [23], which contains audio-visual videos with fine-grained annotations of forged intervals in both modalities. LAV-DF includes lip-sync forgeries, dubbed speech, and partial identity swaps, providing diverse temporal manipulation patterns. Each video is annotated with frame-level start and end times of manipulated segments, enabling supervised training of the boundary module and complementary boundary generator. Following [23], we adopt the official training, validation, and test splits with no overlap in identities or clips across partitions. All temporal localization metrics are reported on the held-out test split.

For clip-level DeepFake detection, we use FaceForensics++ (FF++) [19], a widely used benchmark containing pristine and manipulated videos generated by multiple state-of-the-art facial synthesis methods. Although the manipulations are primarily visual, we still use the synchronized audio stream to train the audio encoder, modality-specific heads, and the cross-modal discrepancy minimization (CMDM) loss. This setting reflects a common real-world scenario in which only one modality is directly manipulated, but cross-modal coherence remains a useful cue. We adopt the standard FF++ split and evaluate on the detection test set, ensuring that identities present in the test videos are not seen during training. Unless otherwise stated, results correspond to the default compression setting used in prior work.

B. Preprocessing and Segment Construction

For each video, we extract face-centred RGB frames at a fixed frame rate and crop them around the facial region of interest. Frames are resized to 96×96 pixels. The corresponding audio is resampled to a fixed sampling rate and converted into log-mel spectrograms with F_m mel-frequency bins and a hop size chosen to maintain approximate alignment with the video frame rate. Training samples are constructed as fixed-length audio-visual segments of T frames. Segments are centred on speech-active regions to maximize informative content and reduce silent or low-motion periods. For LAV-DF, we use the ground-truth temporal annotations to derive frame-wise authenticity labels and manipulation boundaries for supervision of the frame-level heads and boundary module. For FF++, clip-level labels (real/fake) are used for supervising the AVClipHead and CMDM loss.

C. Training Protocol

The model is trained with multi-level supervision on frame-level authenticity, temporal boundaries, and clip-level decisions. The frame-wise logistic regression heads for audio and video are trained using binary cross-entropy to predict per-frame authenticity logits Y_a and Y_v . These logits are both supervised and reused as inputs

to the boundary-aware feature streams X_a and X_v . The boundary module and bidirectional complementary boundary generator are trained to regress manipulation likelihoods over temporal durations up to D frames. For LAV-DF, we decode the ground-truth manipulated intervals into duration-conditioned supervision signals and enforce consistency between forward and backward passes of the complementary generator. The AVClipHead aggregates the fused temporal evidence $\hat{M}_f$ into clip-level scores s_{av} , trained with binary cross-entropy against clip labels. Modality-specific heads h_a and h_v produce s_a and s_v , which are regularized via the CMDM loss to enforce agreement for genuine clips and controlled disagreement for manipulated ones. We optimize all parameters jointly using the Adam optimizer with weight decay and a learning-rate scheduler based on validation performance. Gradient clipping and mixed-precision training are employed to stabilize optimization and reduce memory usage. All ablation experiments are trained under the same schedule, data splits, and augmentation policies to enable fair comparison.

D. Evaluation Metrics

We evaluate the framework on two related tasks: temporal manipulation localization and clip-level deepfake detection. For LAV-DF, we follow a protocol analogous to temporal action detection. The fused boundary representations and refined start/end maps are decoded into manipulation proposals $(t_{\text{start}}, t_{\text{end}})$, which are matched against ground truth using temporal Intersection-over-Union (tIoU). We report (i) mean average precision at tIoU thresholds 0.5 and 0.7, measuring how accurately predicted intervals align with ground truth; (ii) average recall when considering up to 100 proposals per video, indicating how completely the model recovers all forged segments; and (iii) frame-level F1-score at an optimized threshold for start/end boundary predictions, capturing the trade-off between missed and spurious boundary detections. For FF++, we view each video clip as a binary classification instance (real vs. fake) and report the followings metrics. (i) Area Under the ROC Curve (AUC), which summarizes separation between real and fake clips across all thresholds and is robust to class imbalance. Equal Error Rate (EER), i.e., the operating point at which false acceptance and false rejection rates are equal, is relevant for access-control settings. Accuracy and F1 measure thresholded evaluation at the optimal validation threshold, capturing overall classification performance and balance between precision and recall. We additionally perform ablations comparing audio-only, video-only, and fused predictions and evaluate robustness under artificial degradations (motion blur, low-bitrate video compression, and additive noise) to approximate typical consumer capture conditions.

IV. RESULTS AND ANALYSIS

In this section, we present quantitative and qualitative results on FF++ and LAV-DF, analyse the contribution of each architectural component, study robustness under degradation, and assess computational efficiency.

¹ <https://huggingface.co/datasets/ControlNet/LAV-DF>

TABLE III. TEMPORAL MANIPULATION LOCALIZATION RESULTS ON LAV-DF

Model	mAP@0.5	mAP@0.7	AR@100	Boundary F1 (%)
Temporal Convolutional Baseline	61.2	49.5	63.7	64.3
BA-TFD [23] (original)	68.4	55.7	69.5	71.1
BA-TFD+ [23] (improved baseline)	73.2	59.3	72.0	75.0
Proposed (Bidirectional CBG + Uncertainty Fusion)	78.5	64.9	77.8	81.2

TABLE IV. CLIP-LEVEL DEEPFAKE DETECTION PERFORMANCE ON THE FF++ DATASET

Method	AUC (%)	EER (%)	F1 (%)
Xception-DeepFake [29] (visual only)	91.8	10.2	89.4
MesoNet [30] (visual only)	88.5	12.5	86.1
AASIST [31] (audio only)	85.3	14.8	82.0
AV-LipSyncNet [32] (AV baseline)	93.4	8.7	91.2
Proposed (Attention + Uncertainty Fusion)	97.1	5.2	95.6

A. Ablation Study: Contribution of Architecture Modules

To systematically evaluate the contribution of each architectural component, we conduct ablations on both FaceForensics++ (clip-level detection) and LAV-DF (temporal localization). All models are trained under identical conditions. The results are summarized in Table II. Note that the ablation baseline (91.8% AUC) refers to our visual-only encoder without any fusion mechanisms, while BA-TFD+ (95.2% AUC) is a re-implementation of the state-of-the-art transformer-based audio-visual baseline. The ablation study shows a clear and consistent performance gain with each added component. Starting from a baseline AUC of 91.8% and Boundary F1 of 72.5%, attention fusion boosts performance to 94.6% / 75.9%, while uncertainty fusion further improves it to 96.7% / 80.1%. Incorporating bidirectional boundary refinement raises the scores to 96.5% / 79.6%, and the full model with CMDM loss achieves the best results at 97.1% AUC and 81.2% Boundary F1, confirming the effectiveness of each module.

B. Temporal Localization on LAV-DF

Table III reports temporal localization performance on LAV-DF. The temporal convolutional baseline achieves 61.2 mAP@0.5 and a boundary F1 of 64.3%, providing a reference point for simple temporal modeling. The original BA-TFD and improved BA-TFD+ frameworks increase mAP and boundary F1, confirming the benefit of dedicated boundary modeling and audio-visual reasoning. Our full model with bidirectional complementary boundary generator and uncertainty-aware fusion further improves localization quality, reaching 78.5 mAP@0.5, 64.9 mAP@0.7, 77.8 AR@100, and 81.2% Boundary F1. The gains at higher tIoU thresholds (0.7) and the improved boundary F1 indicate that the proposed bidirectional design produces sharper and more temporally accurate boundary estimates. Fig. 4 shows a qualitative example where the baseline model exhibits boundary drift and over-smoothed predictions, whereas the proposed model produces sharp, temporally consistent activations closely aligned with ground-truth forged intervals, particularly at manipulation onsets and offsets.

C. Clip-Level Detection on FaceForensics++

Table IV summarizes clip-level DeepFake detection performance on FF++. Visual-only baselines such as Xception-DeepFake and MesoNet achieve AUCs of 91.8% and 88.5%, respectively, while the audio-only AASIST model reaches 85.3% AUC, confirming that visual artifacts are generally more discriminative than acoustic cues in this dataset. The audio-visual baseline AV-LipSyncNet improves performance to 93.4% AUC and 8.7% EER by exploiting cross-modal synchronization. The proposed model with attention and uncertainty-aware fusion achieves the best overall performance, obtaining 97.1% AUC, 5.2% EER, and 95.6% F1. This corresponds to a substantial reduction in both false

acceptances and false rejections compared to unimodal and fusion baselines. The ROC curves in Fig. 3 further illustrate these gains. The curve for the full model dominates those of prior baselines over most of the false-positive range, indicating more robust separation between real and fake clips across a wide range of operating points.

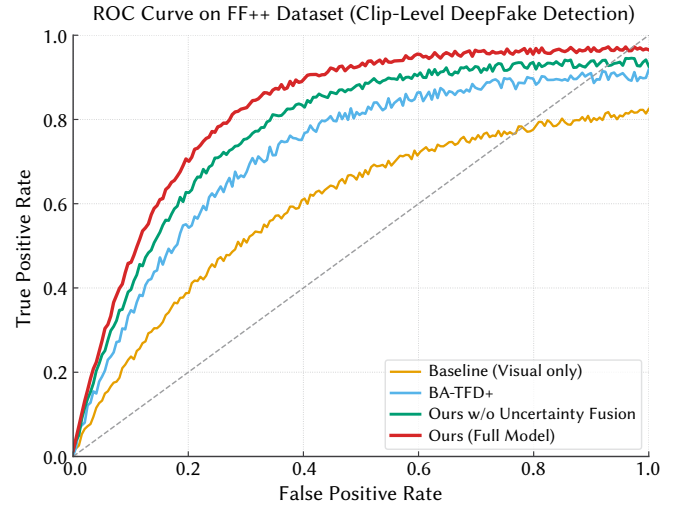


Fig. 3. Receiver Operating Characteristic (ROC) curve comparison on the FaceForensics++ dataset. The proposed full model (red) achieves higher true positive rates across all thresholds, corresponding to an AUC of approximately 0.97, outperforming prior baselines and ablated variants.

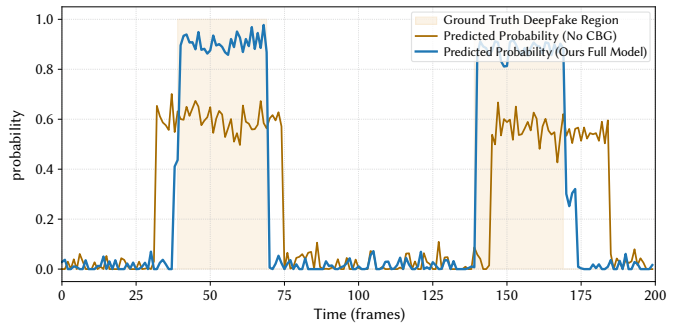


Fig. 4. Temporal localization on LAV-DF. Baseline (orange) shows boundary drift; proposed model (blue) with bidirectional CBG yields sharp predictions aligned with ground-truth (shaded) onsets/offsets.

D. Contribution of Audio and Video Modalities

Table V examines the relative importance of each modality on FF++. The audio-only branch achieves 90.4% AUC and 13.2% EER, while the video-only branch performs better with 95.2% AUC and 7.1% EER, reflecting that FF++ manipulations are primarily visual. Simple late fusion of audio and video (without CMDM) improves AUC to 96.0% and reduces EER to 6.4%, confirming that audio still contributes complementary information. The full model with CMDM achieves the highest AUC (97.1%) and lowest EER (5.2%). This indicates that CMDM not only leverages complementary cues but also regularizes disagreements between modalities in a way that improves decision reliability, especially in challenging or noisy cases.

TABLE V. CONTRIBUTION OF MODALITIES TO CLIP-LEVEL DETECTION

Input Modalities	AUC (%)	EER (%)
Audio only	90.4	13.2
Video only	95.2	7.1
Audio + Video (late fusion, no CMDM)	96.0	6.4
Audio + Video (full model w/ CMDM)	97.1	5.2

E. Boundary Accuracy Vs. Manipulation Duration

Table VI reports boundary F1 on LAV-DF grouped by manipulation duration (short, medium, long). The BA-TFD+ baseline already performs reasonably well but struggles on short manipulations (< 1 s), where boundary cues are subtle and transient. Removing bidirectional CBG moderately improves F1, whereas the full model with bidirectional CBG provides the most consistent gains across all duration ranges. In particular, boundary F1 increases to 78.9%, 82.4%, and 83.2% for short, medium, and long manipulations, respectively. The improvement on short-duration segments indicates that bidirectional reasoning is especially beneficial when manipulation intervals are brief and localized, as it constrains boundary predictions from both temporal directions.

TABLE VI. BOUNDARY F1 ON LAV-DF GROUPED BY MANIPULATION DURATION. SHORT: < 1 s, MEDIUM: 1-3s, LONG: > 3 s

Model	Short	Medium	Long
BA-TFD+ [23]	71.0	75.4	78.1
Ours (No Bidirectional CBG)	72.6	77.1	80.0
Ours (Full Model)	78.9	82.4	83.2

F. Computational Efficiency

Finally, Table VII compares computational cost and latency with a transformer-based BA-TFD+ baseline. The transformer-based model requires 78.4M parameters, 94.2 GFLOPs, and 214 ms per clip on an Nvidia GeForce RTX 4050. By replacing transformer encoders with lightweight 3D CNN (video) and 2D CNN (audio) backbones, our model reduces the parameter count to 21.7M and the computational cost to 18.5 GFLOPs, resulting in an inference latency of 42 ms per clip. A further quantized variant lowers the footprint to 12.9M parameters and 11.2 GFLOPs with latency of 27 ms, while maintaining comparable detection performance. These results

demonstrate that the proposed architecture achieves a favourable trade-off between accuracy and efficiency, delivering state-of-the-art detection and localization performance while being significantly faster and more compact than transformer-based baselines. This supports its suitability for deployment in latency- and resource-constrained consumer scenarios.

TABLE VII. COMPUTATIONAL EFFICIENCY COMPARISON. RUNTIME MEASURED ON AN EMBEDDED-CLASS GPU (GeForce RTX 4050)

Model	Params	GFLOPs	Latency
BA-TFD+ (Transformer Encoders)	78.4M	94.2	214ms
Ours (C3D+CNN Encoders)	21.7M	18.5	42ms
Ours (C3D+CNN + Quantized)	12.9M	11.2	27ms

G. Robustness Under Degradation

Table VIII evaluates clip-level detection robustness on FF++ under controlled degradations: (i) motion blur applied to video frames, (ii) H.264 video compression at 200 kbps, and (iii) additive audio noise at 10 dB SNR. The BA-TFD+ baseline suffers notable drops in AUC (up to ~ 10 points) as degradations become more severe, indicating its sensitivity to visual and acoustic quality. Our model without uncertainty-aware fusion shows smaller performance degradation, suggesting that the lightweight convolutional encoders already confer some robustness. The full model with entropy-based uncertainty fusion is the most stable, with AUC values of 94.2%, 93.5%, and 92.0% under blur, compression, and noise conditions, respectively. Compared to BA-TFD+, the proposed model reduces the degradation by several AUC points across all perturbations, validating the effectiveness of uncertainty-aware modality weighting under realistic capture distortions.

H. Effectiveness of Cross-Modal Discrepancy Minimization

Table IX evaluates how the CMDM objective influences multimodal behaviour during clip-level DeepFake detection. The audio-only classifier (s_a) performs the weakest (AUC 90.4%), reflecting the fact that FF++ contains predominantly visual manipulations. The video-only classifier (s_v) performs substantially better with AUC 95.2%, confirming that visual inconsistencies are the primary cue for detection on this dataset. Simple late fusion of the two classifiers improves performance to AUC 96.0%, indicating that even unmanipulated audio contributes useful complementary information. However, the proposed CMDM loss yields the highest performance at AUC 97.1%, outperforming both individual modalities and naive fusion. The reason is twofold: (i) CMDM enforces prediction alignment for genuine clips, reducing disagreements between modalities and improving calibration; and (ii) for fake clips, CMDM encourages a margin of divergence between the audio and video predictions, which allows the model to detect cross-modal inconsistencies that are typical of AV DeepFakes. This targeted cross-modal regularization ultimately produces a more reliable and discriminative multimodal detector.

TABLE VIII. ANALYSIS ON FF++ UNDER CONTROLLED DEGRADATIONS. VALUES ARE AUC (%) FOR CLIP-LEVEL DEEPFAKE DETECTION

Model	Clean	+Vid. Blur	+Vid. H.264/200kbps	+Audio Noise (10 dB SNR)
BA-TFD+ [23]	95.2	88.7	86.1	84.9
Ours w/o Uncertainty Fusion	96.3	91.5	90.2	88.6
Ours (Full Model)	97.1	94.2	93.5	92.0

TABLE IX. BEHAVIOUR OF MODALITY-SPECIFIC PREDICTIONS UNDER CMDM SUPERVISION ON FF++

Model	AUC (%)	EER (%)
Audio-only head (s_a)	90.4	13.2
Video-only head (s_v)	95.2	7.1
Late fusion (no CMDM)	96.0	6.4
Full model with CMDM (s_{av})	97.1	5.2

V. DISCUSSION

The results on FF++ and LAV-DF demonstrate that the proposed framework simultaneously advances three core aspects of AV DeepFake detection: (i) accuracy in clip-level classification, (ii) temporal precision in manipulation localization, and (iii) computational efficiency suitable for on-device deployment. The improvement from 91.8% AUC in the baseline configuration to 97.1% AUC in the full model, together with an 81.2% Boundary F1 score on LAV-DF, indicates that the proposed design does more than simply combine audio and visual cues—it leverages their interactions in a structured and uncertainty-aware way. The bidirectional complementary boundary generator provides sharper and temporally more consistent boundaries, particularly for short and medium-duration forgeries, while the entropy-based fusion scheme stabilizes decisions when one of the modalities is degraded.

A key outcome of our analysis is that not all multimodal fusion strategies are equally beneficial. Simple late fusion of modality-specific decisions already improves over unimodal baselines, but the gains plateau quickly. In contrast, attention-based fusion operating on boundary maps and feature representations, followed by uncertainty-aware re-weighting, yields consistent improvements in both mAP and AUC. This suggests that DeepFake detection benefits from treating audio and video not as independent experts, but as correlated sources whose agreement and disagreement patterns carry diagnostic value. The Cross-Modal Discrepancy Minimization (CMDM) objective further reinforces this behaviour by explicitly rewarding agreement on genuine clips and controlled divergence on manipulated clips, thereby aligning the fused prediction with the underlying audio-visual coherence structure. Another important finding concerns robustness under realistic degradations. The entropy-driven fusion gate enables the model to adaptively prioritize the more reliable modality when faced with blur, heavy compression, or noisy audio, leading to markedly smaller performance drops compared with transformer-based baselines. This is particularly relevant for consumer environments, where capture hardware, bandwidth constraints, and background conditions cannot be controlled. The observed robustness implies that model decisions are not overly tied to brittle artifacts or high-fidelity conditions; instead, the system maintains discriminative power when only partial or noisy evidence is available.

At the same time, the analysis highlights several limitations that motivate future work. First, both FF++ and LAV-DF, while widely used, capture a limited range of manipulation types, recording conditions, languages, and demographic attributes. There is a risk that models optimized on these datasets may underperform on emerging DeepFake pipelines, cross-lingual scenarios, or user-generated content with more diverse speaking styles and camera viewpoints. Second, our approach focuses on short segments centred around speech-active regions; extending the framework to longer, unconstrained videos with complex conversational dynamics and multiple speakers remains an open challenge. Third, although our method is computationally efficient compared with transformer-based models, further exploration of hardware-aware design (e.g., platform-specific quantization and pruning) is needed to fully exploit resource-limited edge devices. Finally, broader deployment of deepfake detection systems raises

ethical and societal considerations. While stronger detectors can help mitigate misuse of synthetic media, widespread automated screening also introduces questions about privacy, transparency, and potential biases against certain speakers, accents, or visual appearances. Our framework provides interpretable intermediate outputs (e.g., boundary maps and modality-specific scores) that can support explainability, but more work is required on user-facing interfaces, calibration across populations, and safeguards against adversarial adaptation. Addressing these aspects will be crucial for building trustworthy AV DeepFake detection tools that can be responsibly integrated into real-world platforms.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Despite the strong performance of the proposed AV-DeepFake-Net framework, several limitations should be acknowledged. The model is trained and evaluated on benchmark datasets such as LAV-DF and FaceForensics++, which, although widely used, may not fully capture the diversity of real-world scenarios. These datasets exhibit biases in terms of demographic representation, recording conditions, and types of manipulations. As a result, the model may experience performance degradation when applied to data from different domains, such as videos captured with varying camera qualities, lighting conditions, or cultural contexts. Addressing domain shift remains an important challenge for real-world deployment. The proposed framework relies on audio-visual consistency cues, including speech patterns and lip synchronization. However, speech characteristics vary significantly across languages, accents, and speaking styles. Since the training data is primarily based on limited linguistic distributions, the model may not generalize optimally to unseen languages or multilingual environments. Similarly, domain variations such as background noise, recording devices, and environmental conditions can affect the reliability of learned representations. DeepFake generation methods are rapidly evolving, introducing new manipulation strategies that may not be represented in current datasets. While the proposed model demonstrates robustness to common distortions such as noise and compression, its performance against entirely novel or adversarial manipulation techniques remains uncertain. In particular, advanced generative models that produce highly consistent audio-visual signals may reduce detectable discrepancies, posing challenges for current detection approaches.

Future work will focus on improving generalization through domain adaptation techniques, incorporating more diverse and multilingual datasets, and developing self-supervised or continual learning strategies to better handle unseen manipulation techniques. Additionally, integrating explainability mechanisms may further enhance trust and transparency in real-world deployment.

VII. CONCLUSION

This paper presented an attention guided uncertainty aware audio-visual framework for deepfake detection that is explicitly designed with on-device deployment in mind. By combining lightweight 3D convolutional encoders for video and 2D convolutional encoders for audio with boundary-centric temporal modelling, the proposed system jointly captures spatial, spectral, prosodic, and temporal cues indicative of manipulation. The bidirectional complementary boundary generator yields sharp and temporally consistent localization of forged segments, while the entropy-based uncertainty fusion gate dynamically downweights unreliable modalities under blur, compression, or noise. On top of these components, the AVClipHead and Cross-Modal Discrepancy Minimization (CMDM) objective enforce coherent decision-making across modalities, encouraging agreement

for genuine clips and informative discrepancies for manipulated ones. Extensive experiments on LAV-DF and FaceForensics++ show that the framework delivers both strong accuracy and practical efficiency. On FF++, the model achieves 97.1% AUC and 95.6% F1 for clip-level detection, reducing the equal error rate to 5.2% and outperforming unimodal and existing multimodal baselines. On LAV-DF, it attains a Boundary F1 of 81.2% and higher mAP across tIoU thresholds, demonstrating precise temporal localization of forged intervals. Robustness evaluations under motion blur, aggressive H.264 compression, and noisy audio further confirm that uncertainty-aware fusion substantially mitigates performance degradation compared with transformer-based architectures. At the same time, the use of compact convolutional backbones and boundary-focused reasoning yields up to a $5\times$ reduction in inference latency, supporting real-time processing on embedded and mobile platforms.

CREDiT AUTHORSHIP CONTRIBUTION STATEMENT

Nasir Saleem: Conceptualization, Methodology, Software, Investigation, Writing – Original Draft.

Ahmad Ali: Methodology, Software, Validation, Formal Analysis, Writing – Review & Editing.

Zhuoqi Zeng: Resources, Funding Acquisition, Writing – Review & Editing.

Adeel Hussain: Investigation, Validation, Formal Analysis.

Sami Bourouis: Methodology, Supervision, Writing – Review & Editing.

Sami Dhahbi: Validation, Visualization, Writing – Review.

Afef Dhahbi: Data Curation, Formal Analysis.

DECLARATION OF CONFLICTS OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are publicly available in the LAV-DF dataset at <https://huggingface.co/datasets/ControlNet/LAV-DF>.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/601/46. This work was supported also by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R928), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

- [1] M. Talha Usman, H. Khan, S. Kumar Singh, M. Young Lee, J. Koo, "Efficient Deepfake Detection via Layer-Frozen Assisted Dual Attention Network for Consumer Imaging Devices," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 1, pp. 281–291, 2025, doi: <https://doi.org/10.1109/TCE.2024.3476477>
- [2] G. Zhang, Y. Liang, K. Tian, J. Yi, H. Alsolai, M. Liu, X. Hu, "Leveraging Spatial-Temporal Illumination Features and Convolution-Transformer Hybrid Networks for Deepfake Video Detection," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2025, doi: <https://doi.org/10.1109/TCE.2025.3624764>
- [3] G. Zhang, Q. Li, M. Gao, S. Guo, G. Jeon, "Detecting Sequential Deepfake Manipulation via Spectral Transformer With Pyramid Attention in Consumer IoT," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 6371–6380, 2025, doi: <https://doi.org/10.1109/TCE.2024.3414319>
- [4] H. B. Eutamene, M. Keita, W. Hamidouche, A. Taleb-Ahmed, H. Liz-López, A. Martín, D. Camacho, A. Hadid, "Advances in AI-Generated Images and Videos," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 173–208, 2024.
- [5] R. Mahum, A. Irtaza, A. Javed, "EDL-Det: A robust TTS synthesis detector using VGG19-based YAMNet and ensemble learning block," *IEEE Access*, vol. 11, pp. 134701–134716, 2023.
- [6] Z. Yu, R. Cai, Z. Li, W. Yang, J. Shi, A. C. Kot, "Benchmarking joint face spoofing and forgery detection with visual and physiological cues," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4327–4342, 2024.
- [7] S. Waseem, S. A. R. S. A. Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa, M. E. E. Dalam, "DeepFake on face and expression swap: A review," *IEEE Access*, vol. 11, pp. 117865–117906, 2023.
- [8] A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, A. C. Kot, "Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 1168–1182, 2023.
- [9] C. Kong, B. Chen, H. Li, S. Wang, A. Rocha, S. Kwong, "Detect and locate: Exposing face manipulation by semantic-and noise-level telltales," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1741–1756, 2022.
- [10] A. Raza, K. Munir, M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, p. 9820, 2022.
- [11] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, others, "The face deepfake detection challenge," *Journal of Imaging*, vol. 8, no. 10, p. 263, 2022.
- [12] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, others, "Deepfakes detection with automatic face weighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 668–669.
- [13] H. Khalid, M. Kim, S. Tariq, S. S. Woo, "Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors," in *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, 2021, pp. 7–15.
- [14] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019.
- [15] D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, S. Tubaro, "A robust approach to multimodal deepfake detection," *Journal of Imaging*, vol. 9, no. 6, p. 122, 2023.
- [16] A. Anshul, S. Gopal, D. Rajan, E. S. Chng, "Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 13826–13836.
- [17] Y. Wang, Q. Sun, J. Zhang, D. Rong, C. Shen, X. Wang, "Improving deepfake detection with predictive inter-modal alignment and feature reconstruction in audio-visual asynchrony scenarios," *Information Fusion*, p. 103708, 2025.
- [18] Y. Wang, Y. Wang, Q. Zhang, H. Nishizaki, M. Li, "VCAPAV: A Video-Caption Based Audio-Visual Deepfake Detection Dataset," in *Proc. Interspeech 2025*, 2025, pp. 3908–3912.
- [19] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.
- [20] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213.
- [21] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.
- [22] H. Khalid, S. Tariq, M. Kim, S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.
- [23] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, M. Hayat, "Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization," *Computer Vision and Image Understanding*,

vol. 236, p. 103818, 2023.

- [24] S. K. Datta, T. Ranga, C. Sun, S. Lyu, "PIA: Deepfake Detection Using Phoneme-Temporal and Identity-Dynamic Analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1596–1606.
- [25] Z. Li, Z. Teng, B. Zhang, J. Fan, "Open-Unfairness Adversarial Mitigation for Generalized Deepfake Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 698–707.
- [26] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.
- [27] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, others, "Df40: Toward next-generation deepfake detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 29387–29434, 2024.
- [28] A. H. Abdelaziz, others, "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition," in *Interspeech*, 2017, pp. 3752–3756.
- [29] D. A. Alkurdi, M. Cevik, A. Akgundogdu, "Advancing Deepfake Detection Using Xception Architecture: A Robust Approach for Safeguarding against Fabricated News on Social Media," *Computers, Materials & Continua*, vol. 81, no. 3, 2024.
- [30] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*, 2018, pp. 1–7, IEEE.
- [31] K. Borodin, V. Kudryavtsev, D. Korzh, A. Efimenko, G. Mkrtchian, M. Gorodnichev, O. Y. Rogov, "AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 Challenge," in *Proc. ASVspoof 2024*, 2024, pp. 48–55.
- [32] S. A. Jeevakumari, K. Dey, "LipSyncNet: a novel deep learning approach for visual speech recognition in audio-challenged situations," *IEEE Access*, vol. 12, pp. 110891–110904, 2024.



Nasir Saleem

Nasir Saleem received Ph.D. in Speech Processing and Deep Learning from UET Peshawar in 2021, working in BIUH Hainan China. He was a Research Fellow at ENU, UK and was a postdoctoral Fellow at IIUM, working on AI-based speech processing algorithms. Previously, he served as an assistant Professor and is now an Associate Professor at Gomal University. His research interests

include Human–Machine Interaction, Speech Enhancement, AV Processing, and Machine Learning.



Ahmad Ali

Ahmad Ali received the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, China, where he focused on advanced machine learning and intelligent systems. He was a Post-Doctoral Researcher at Shenzhen University, China. Currently he is doctoral lecturer in Hainan Bielefeld University of Applied Sciences, Hainan, China.



Zeng Zhuoqi

Zeng Zhuoqi is a distinguished professor and academic leader in the fields of computer science and technology, as well as economic engineering. He currently serves as a subject leader at Hainan Bieke University. Recognized as a national-level young talent, Professor Zeng has made significant contributions to interdisciplinary innovation and academic development. He is also actively engaged

in international academic and professional communities, serving as a council member and specially appointed expert of the German Mechanical and Electrical Engineering Society.



Adeel Hussain

Adeel Hussain obtained his BSc in Audiology from De Montfort University and has over a decade of experience as a practicing Clinical Audiologist, working across both the NHS and private sectors. He is currently pursuing a PhD at Edinburgh Napier University, where his research focuses on developing an audio-visual testing framework for current and future multimodal hearing technologies. His

work draws on extensive experience with a wide range of hearing technologies, including implantable devices such as cochlear implants and bone-anchored hearing aids.



Sami Bourouis

Sami Bourouis received B.Sc, M.Sc., and Ph.D. degrees in computer science from the University of Tunis, Tunisia, in 2003, 2005, and 2011, respectively. He is currently a Professor at the College of Computers and Information Technology, Taif University, Saudi Arabia. His research interests include data mining, image processing, AI, machine learning, cyber security, and pattern recognition.



Sami Dhahbi

Sami Dhahbi received the engineering and M.S. degrees from the National School of Computer Science, University of Manouba, Tunisia, in 2005 and 2006, respectively, and the Ph.D. degree in computer science from the University of Tunis, El Manar, Tunisia, in 2016. He is currently an assistant professor of computer science at King Khalid University, Saudi Arabia. He is also a member of the

LIMITIC Research Laboratory at the University of Tunis, El Manar.

Afef Dhahbi

Afef Dhahbi received the Engineer, M.Sc., and Ph.D. degrees in computer science. Her research interests include data mining, AI, machine learning, cyber security, and pattern recognition.