

Article

Imbalance-Aware Cross-Modal Focal Modulation for Cross-Dataset Audio-Visual Deepfake Detection

Shahad Mohammad Bn Dokiey ¹, Tariq M. Khan ^{1,2,*}  and Qazi Emad Ul Haq ^{1,2} 

¹ Department of Cybersecurity and Digital Forensics, College of Forensic & Investigative Sciences, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia; 202400256@student.nauss.edu.sa

² Center of Artificial Intelligence for Security, Naif Arab University for Security Sciences, Riyadh 14812, Saudi Arabia; qabdulrab@nauss.edu.sa

* Correspondence: tkhan@nauss.edu.sa; tariq045@gmail.com

Abstract

Audio-visual deepfake detection remains challenging under cross-dataset distribution shift, especially when the source-domain training data are severely imbalanced. Existing middle-fusion detectors often rely on softmax-based cross-attention, which can learn sharp source-domain token interactions and may transfer poorly to unseen datasets. This study proposes FocalNet, an audio-visual detector that replaces the cross-attention block of the 2D3MF framework with cross-modal focal modulation. The proposed module aggregates multi-scale temporal context before audio-visual interaction, enabling softmax-free contextual modulation between visual MARLIN features and audio EAT features. We evaluate the method under a strict FakeAVCeleb-to-DFDC protocol, where all training and validation is performed on FakeAVCeleb and the DFDC is used only as an unseen target-domain test set. Compared with the reproduced 2D3MF baseline, which collapses to a single-class prediction pattern on the DFDC, FocalNet achieves substantially stronger zero-shot score separation, with a DFDC ROC-AUC of 0.9324. Thresholded analysis further shows an improved balanced accuracy, macro F1 score, and MCC when the frozen source-domain operating point is applied. The model also preserves practical efficiency, requiring comparable FLOPs and lower per-sample inference time than the reproduced baseline. These findings suggest that cross-modal focal modulation is a promising alternative to attention-based middle fusion for audio-visual deepfake detection under dataset shifts, while broader validation across additional unseen datasets, multi-seed training, and deployment-oriented calibration remain important for future work.

Keywords: audio-visual deepfake detection; cross-dataset generalization; focal modulation; multimodal fusion; class imbalance; confusion matrix; explainable artificial intelligence; FakeAVCeleb; DFDC



Academic Editor: Luis Javier Garcia Villalba

Received: 4 June 2026

Revised: 15 July 2026

Accepted: 17 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recent advances in generative adversarial networks (GANs), diffusion models, and neural vocoders have enabled high-quality deepfake media in which facial appearance and speech can be manipulated simultaneously [1,2]. These technologies are increasingly accessible, which has increased the risk of real-world misuse and raised concerns about the integrity of digital media, public trust, and individual safety.

Although substantial progress has been made in unimodal deepfake detection, including methods that target visual artifacts [3] or audio anomalies [4], audio-visual deepfake

detection remains comparatively less explored. Multimodal detection is attractive because genuine media usually exhibit strong cross-modal consistency; speech, facial motion, lip movement, and temporal dynamics are naturally aligned. Manipulated media, on the contrary, can introduce subtle inconsistencies between modalities [5].

A key limitation of many existing approaches is the evaluation protocol. Most models are assessed in dataset settings, where training and test samples originate from the same distribution. Although such protocols may report high performance, sometimes exceeding an AUC of 0.99, they do not fully reflect real-world deployment, where detectors must generalize to unseen datasets, identities, compression pipelines, manipulation methods, and acquisition conditions. In practice, a model trained on a dataset, such as FakeAVCeleb, can perform poorly on another dataset, such as the Deepfake Detection Challenge (DFDC), revealing a substantial generalization gap [6,7].

One factor that may contribute to this gap is the use of softmax-based cross-attention in middle-fusion architectures such as 2D3MF [8]. Softmax-normalized pairwise attention can amplify small differences in token similarity scores, producing sharp source-domain interactions. Moreover, conventional attention primarily models global token relationships and may not sufficiently emphasize local temporal artifacts that characterize audio-visual manipulation.

To address these limitations, this paper proposes a focal modulation network (FocalNet) for audio-visual deepfake detection. The proposed module replaces the cross-attention block in 2D3MF with cross-modal focal modulation [9], which aggregates multi-scale temporal context before audio-visual interaction. This first-aggregate, last-interact design enables the model to capture both local temporal variations and broader structural cues without relying on token-to-token cross-attention. Because additional experiments show that naive fully independent sigmoid focal gating and naive class balancing can degrade transfer, the revised manuscript treats these variants as diagnostic ablations and retains the original FocalNet configuration as the final model.

The need for FocalNet follows from a sharper gap in current audio-visual deepfake detection. Source-domain cross-attention can provide strong in-dataset fusion, but under severe imbalance, it may also amplify dataset-specific audio-visual correspondences instead of learning transferable manipulation evidence. A softmax-free, multi-scale modulation mechanism is therefore needed to aggregate local and broader temporal context before cross-modal interaction. The present study tests this idea under a strict FakeAVCeleb-to-DFDC setting and evaluates it with imbalance-aware metrics rather than accuracy alone.

The main contributions of this work are as follows:

1. We introduce a FocalNet-based audio-visual deepfake detector that replaces the softmax-based cross-attention block of 2D3MF with cross-modal focal modulation and evaluates it under a strict FakeAVCeleb $\rightarrow$ DFDC protocol.
2. We provide an imbalance-aware evaluation with exact split counts, confusion matrices, class-wise precision, recall, and F1 scores, balanced accuracy, MCC, EER, and threshold-adjusted analysis, showing that accuracy alone can hide complete failure on the minority real class.
3. We verify the reproduced 2D3MF baseline and show that, under the strict FakeAVCeleb-only training setting, it collapses to an all-fake prediction pattern on the DFDC, which establishes a transparent and reproducible baseline for cross-dataset analysis.
4. We conduct controlled optimizer/loss ablations showing that SGDM and focal loss alter the class-recall trade-off, while the strongest cross-dataset separability is obtained by the focal modulation architecture trained with Adam and BCE.
5. We add three diagnostic architecture and imbalance ablations: an independent sigmoid-gated focal variant, a balanced-sampler variant, and a balanced-sampler

variant with residual modality reliability gating. These experiments clarify why naive gating or balancing does not replace the original FocalNet configuration.

6. We develop a cross-modal explainability analysis based on Integrated Gradients, limited to feature-level temporal and modality-level attribution patterns without anatomical region or pixel-level image mapping claims.

2. Related Work

This section reviews the work most relevant to the proposed audio-visual deepfake detector. Rather than surveying the entire deepfake detection literature, the discussion focuses on four themes that directly motivate this study: unimodal visual and audio representations, audio-visual fusion, cross-dataset generalization under imbalance, and explainability for multimodal detection.

2.1. Visual and Audio Deepfake Detection

Early visual deepfake detectors mainly relied on frame-level appearance artifacts, but video-based methods became more effective, as manipulation traces often appear as temporal inconsistencies in facial motion, audio-visual synchronization, and local texture patterns. Spatiotemporal CNNs and pseudo-3D networks were introduced to jointly model appearance and motion, showing that temporal information can improve recognition beyond image-level analysis [10,11]. Transformer-based video models later extended this direction by using self-attention to capture long-range dependencies between frames [12]. However, these models are often data-hungry and may overfit to dataset-specific artifacts when trained under limited or biased deepfake datasets. Recent work has also explored real-time detection using gaze patterns, blink analysis, lip movement synchronization, and hybrid deep learning approaches [13–17], highlighting the growing importance of efficient and interpretable detection pipelines.

Recent facial video representation learning has therefore shifted toward self-supervised and domain-aware pretraining. Masked autoencoding methods, such as VideoMAE, learn transferable spatiotemporal features by reconstructing missing regions from video sequences [18,19]. For face-centered videos, MARLIN further incorporates region-guided masking during pretraining, encouraging the representation to encode face-related video structures [20]. In the present work, MARLIN is used as a compact visual feature extractor, but its embedding channels are not treated as anatomically ordered facial region variables. Accordingly, the present work uses MARLIN features as the visual input representation.

Audio deepfake detection has followed a similar transition from handcrafted acoustic descriptors to learned representations. Traditional approaches based on MFCC or filter-bank features can capture spectral distortions in synthetic speech, but they have a limited ability to represent long-range prosody, speaker dynamics, and semantic-acoustic consistency [21,22]. Self-supervised speech models such as wav2vec 2.0, HuBERT, and data2vec improved the learning of audio representation by exploiting large-scale unlabeled audio and masked prediction objectives [23–25]. More recent audio transformers and masked spectrogram models further improve temporal modeling and scalability [26,27]. In this work, Efficient Audio Transformer (EAT) features are adopted because they provide compact and informative temporal representations for the audio stream [28].

Although unimodal visual and audio detectors can perform well in controlled settings, they remain vulnerable to domain shift. Visual models may learn compression patterns, dataset-specific identities, or manipulation artifacts that do not transfer to other datasets, while audio models may be affected by the recording quality, noise, vocoder type, and segment length [4,29,30]. These limitations motivate audio-visual detection, where

the model can exploit both modality-specific artifacts and inconsistencies between facial motion and speech.

2.2. Audio-Visual Deepfake Detection and Fusion

Audio-visual deepfake detection is based on the observation that authentic videos usually preserve natural correspondence between speech and facial behavior, whereas manipulated media may introduce cross-modal inconsistencies. Early audio-visual methods formulated this problem as a task of synchronization or dissonance, learning whether audio and facial motion are mutually consistent [5,31]. Emotion-based methods extended this idea by detecting mismatches between facial expression, voice emotion, and semantic content, showing that affective cues can provide useful evidence beyond low-level artifacts [32–34].

The fusion strategy is central to multimodal detection. Feature-level fusion combines audio and visual embeddings before classification, enabling the model to learn joint representations but also making it susceptible to modality dominance. Decision-level fusion combines independent modality predictions, which are robust to missing modalities but cannot model fine cross-modal interactions [35]. More recent methods use cross-modal fusion, attention, contrastive learning, complementary masking, or dynamic weighting to explicitly model inter-modal relationships [36–38]. These methods improve detection by capturing interactions that are not visible from either modality alone.

The 2D3MF framework is particularly relevant because it performs middle fusion between audio and visual features through bidirectional cross-attention [8]. Although cross-attention can model dependencies between modalities, it relies on softmax-normalized token-to-token attention weights. In cross-dataset settings, this may encourage sharp and overconfident token interactions that fit the source dataset but fail to transfer to unseen datasets. The proposed method maintains the middle-fusion principle of 2D3MF but replaces the pairwise cross-attention operation with a focal modulation block. Focal modulation aggregates the multi-scale temporal context before feature interaction and performs contextual modulation through lightweight gating [9]. This design is expected to better preserve local temporal artifacts and reduce dataset-specific attention behavior.

In comparative terms, prior audio-visual deepfake methods differ from the present work in both fusion mechanism and evaluation protocol. Synchronization- and dissonance-based approaches mainly examine whether speech and facial motion are mutually consistent, while emotion-based methods focus on affective disagreement between voice, expression, and semantic content. More recent fusion-based detectors use cross-attention, contrastive learning, complementary masking, or dynamic modality weighting; however, many reported results were obtained under in-dataset or mixed-dataset protocols in which the target dataset contributed to training. In contrast, the present work focuses on strict FakeAVCeleb-to-DFDC zero-shot transfer, replaces softmax-based middle fusion with cross-modal focal modulation, and emphasizes imbalance-aware evaluation through the confusion matrices, class-wise recall, balanced accuracy, MCC, EER, PR-AUC, Brier score, and operating-point sensitivity.

2.3. Generalization, Class Imbalance, and Evaluation Protocols

A recurring limitation in deepfake detection research is that strong in-dataset performance does not necessarily imply real-world robustness. Many detectors are trained and tested on splits from the same dataset, where manipulation methods, identities, compression settings, and acquisition conditions are shared. Under such protocols, models may achieve an extremely high AUC but fail when evaluated on a different benchmark. Cross-dataset evaluation is therefore a more realistic test of generalization because the training and testing distributions differ. Datasets such as FakeAVCeleb, Celeb-DF, Face-

Forensics++, and the DFDC vary substantially in manipulation type, audio quality, visual quality, and class composition [3,6,7,39].

Class imbalance further complicates evaluation. In binary deepfake detection, a detector trained on a heavily skewed dataset may favor the majority class and still report high accuracy, while minority-class recall remains poor. This issue is especially important in FakeAVCeleb, where the filtered training data used in this work are strongly skewed toward fake samples. Metrics such as the AUC, macro F1 score, class-wise recall, and EER are therefore more informative than accuracy alone [40,41]. Previous audio-visual methods often report results under in-dataset or mixed-dataset training protocols, sometimes including the target dataset during training [37,38,42]. In contrast, this paper adopts a strict cross-dataset protocol: training is performed only on FakeAVCeleb, and evaluation is conducted on the unseen DFDC dataset. This setting directly measures whether the learned audio-visual representation transfers under distribution shift.

2.4. Explainability in Deepfake Detection

Explainability is important for forensic deployment because detection results must be interpretable, auditable, and connected to meaningful manipulation evidence. Attribution methods such as Integrated Gradients estimate the contribution of input features to a model prediction and can reveal whether a detector relies on plausible cues or spurious correlations [43]. In video deepfake detection, feature-level explanations can be examined across temporal segments and modalities, helping to identify whether decisions are driven by visual, audio, or cross-modal evidence. For audio-visual models, explainability can also indicate how the detector balances modalities across real and fake samples.

Most existing audio-visual deepfake detectors emphasize predictive performance, while relatively fewer studies analyze cross-modal attribution behavior under unseen test distributions. This paper addresses that gap by applying Integrated Gradients to the learned audio and visual feature sequences and summarizing attributions temporally and by modality. Because the analysis is performed on pre-extracted MARLIN and EAT features, it is used only as a post hoc feature-level diagnostic; it is not treated as evidence of anatomical region, pixel-level, or causal localization.

2.5. Research Gap

The reviewed literature shows that audio-visual fusion is promising for deepfake detection, but three gaps remain. First, many methods are evaluated under in-dataset or mixed-dataset settings, making it difficult to assess transfer to unseen domains. Second, cross-attention-based fusion may learn sharp source-domain interactions that do not generalize well under distribution shifts. Third, limited attention has been paid to explaining how multimodal detectors distribute importance across time and modalities under cross-dataset evaluation. To address these gaps, this work proposes a cross-modal focal modulation fusion module, evaluates it under a strict FakeAVCeleb-to-DFDC protocol, performs additional diagnostic ablations for gating, sampling, and modality reliability, and analyzes its predictions using Integrated Gradients-based cross-modal explainability.

3. Methodology

This section describes the experimental pipeline used in this study. As shown in Figure 1, the pipeline consists of three stages. First, we reproduce the 2D3MF baseline under the same feature-level audio-visual setting. Second, we replace the baseline cross-attention module with the proposed cross-modal focal modulation block. Third, we perform post hoc explainability analysis using Integrated Gradients, limited to feature-level temporal

attribution and modality-level attribution. The cross-dataset protocol trains the models on FakeAVCeleb and evaluates them as zero-shot on the DFDC.

Proposed Framework for Cross-Dataset Audio-Visual Deepfake Detection

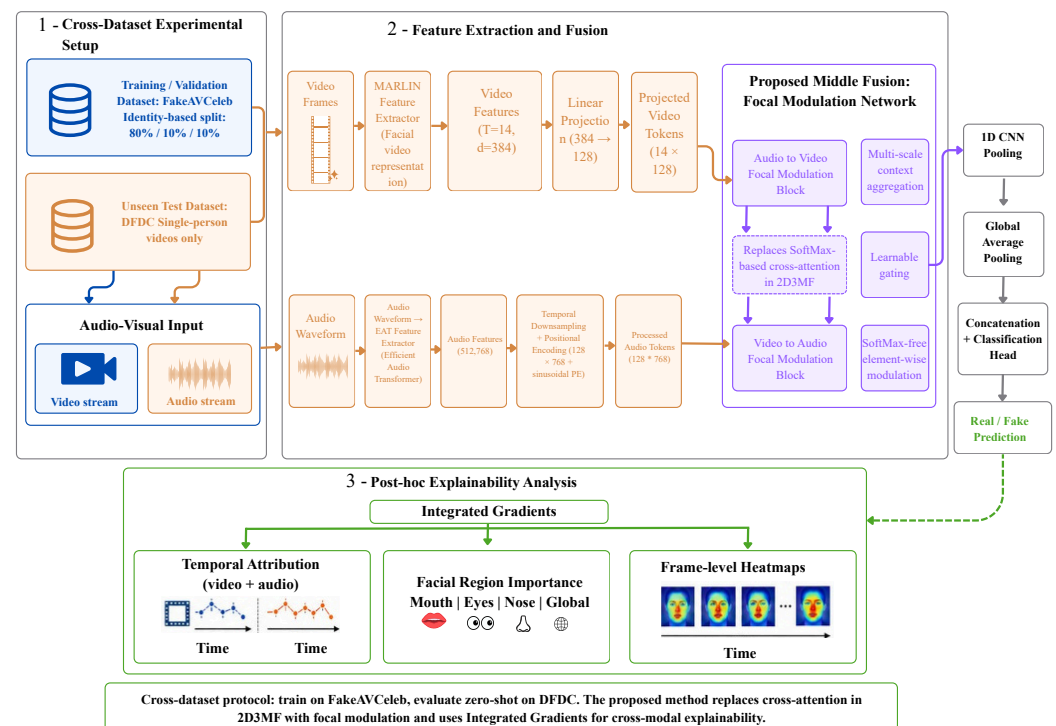


Figure 1. Overall architecture of the proposed FocalNet-based audio-visual deepfake detection framework. Blue identifies source-domain and video-input elements; orange identifies dataset and feature-processing elements; purple identifies the proposed cross-modal focal modulation blocks; gray identifies pooling and classification components; and green identifies prediction and post hoc explainability outputs. All schematic elements were redrawn by the authors for this manuscript; any generic icons used in the schematic are author-created or open-license graphical elements and do not reproduce copyrighted third-party artwork.

3.1. Experimental Set-up

3.1.1. Datasets

The FakeAVCeleb [6] dataset served as the source-domain training dataset. The data include four categories: FakeVideo-FakeAudio (FVFA), RealVideo-FakeAudio (RVFA), FakeVideo-RealAudio (FVRA), and RealVideo-RealAudio (RVRA), where RVRA is treated as the genuine class and the remaining categories are treated as fake. To reduce identity leakage, the original split protocol placed all videos from the same identity into a single split. The final feature-level splits used in this work were verified directly from the split files, and they are reported in Table 1. The resulting FakeAVCeleb training set was extremely imbalanced, with 400 real samples and 16,715 fake samples, corresponding to a 41.8:1 fake-to-real ratio. The validation and test splits were similarly skewed, with fake-to-real ratios of 43.1:1 and 43.5:1, respectively.

The Deepfake Detection Challenge (DFDC) dataset [7] was used only as an unseen target-domain benchmark. Following the audio-visual setting of the baseline framework, only videos containing a single visible person were retained. The final DFDC test subset contained 5200 samples, with 877 real and 4323 fake samples. No DFDC sample was used during training, validation, threshold selection, early stopping, or hyperparameter selection.

Table 1. Verified dataset composition used in the experiments. Counts were obtained from the final split files used by the code.

Dataset	Split	Real	Fake	Total	Fake:Real
FakeAVCeleb	Train	400	16,715	17,115	41.8:1
FakeAVCeleb	Validation	50	2154	2204	43.1:1
FakeAVCeleb	Test	50	2175	2225	43.5:1
FakeAVCeleb	Total	500	21,044	21,544	42.1:1
DFDC	Test	877	4323	5200	4.9:1

3.1.2. Feature Extraction

Following the 2D3MF pipeline, all subsequent experiments used pre-extracted features rather than raw video frames, which substantially reduced the computational cost while preserving informative multimodal representations. For the visual modality, we extracted MARLIN features [20]. MARLIN is a facial video representation model based on a masked autoencoder architecture and is pretrained on YouTube face videos using region-guided masking. This training strategy encourages the representation to capture face-related temporal structures, making MARLIN suitable as a compact visual feature source for deepfake detection. The extracted visual features contained 14 temporal segments, each with 384 dimensions.

For the audio modality, we used Efficient Audio Transformer (EAT) features [28]. The EAT provides temporal audio representations that can capture speech-related inconsistencies relevant to audio-visual manipulation. The extracted audio features contained 512 temporal frames, each with 768 dimensions.

3.1.3. Linear Projection and Positional Encoding

Before performing cross-modal fusion, we employed targeted preprocessing steps that were performed independently in each modality to align the feature representations and enable proper interaction between the two streams. For the visual stream, the MARLIN features with a size of 384 were linearly projected to a hidden size of 128 using a learnable linear projection. This reduction in dimensions helped reduce computational complexity and ensured alignment of the feature dimensionality with the internal representation space of the fusion module.

No dimensionality reduction was applied along the audio feature channel dimension. Along the temporal axis, the 512 EAT frames were reduced to 128 frames using temporal pooling in the feature alignment stage, yielding audio features 128×768 in size. Sinusoidal positional encodings were then added to preserve temporal ordering. After these steps, the dimensions of the video features became 14×128 , while those of the audio features were 128×768 .

3.1.4. Training Protocol

Unless stated otherwise, model optimization was performed using the Adam algorithm with an initial learning rate of 1×10^{-4} . The exponential decay rates for the first- and second-moment estimates were set to $\beta_1 = 0.5$ and $\beta_2 = 0.9$, respectively, rather than the common defaults of 0.9 and 0.999. The lower β_1 reduced the influence of accumulated gradient history, which was found empirically on the validation set to stabilize training under the severe class imbalance of FakeAVCeleb. A ReduceLROnPlateau scheduler reduced the learning rate by a factor of 0.5 when no improvement in validation loss was observed for seven consecutive epochs, with a minimum learning rate of 1×10^{-8} .

Training took place for up to 300 epochs. The validation AUC score served as the primary criterion for selecting the best-performing checkpoint. The batch size chosen was

16 for all experiments. Mixed-precision training (FP16) was performed using PyTorch Lightning version 2.4.0 to reduce memory usage and improve the computing speed. Gradient clipping with a maximum norm of 1.0 was implemented to enhance the stability of optimization during the early phases of training. All experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU. A fixed random seed of 42 was used to improve reproducibility.

In the main proposed configuration, BCEWithLogitsLoss was used because it combines a sigmoid layer and binary cross-entropy in a single numerically stable operation suited to mixed-precision training. This is the default configuration; in the controlled optimizer/loss ablation, the architecture is held fixed, and the optimizer is varied between Adam and SGDM, while the loss is varied between BCEWithLogitsLoss and focal loss. This design tested whether the reported gains arose only from the optimizer or imbalance-aware loss or from the focal modulation architecture itself.

3.1.5. Evaluation Metrics and Thresholding

Because the FakeAVCeleb splits were severely imbalanced, accuracy and ROC-AUC are not sufficient on their own. We therefore report the accuracy (ACC), balanced accuracy (BACC), macro F1 score, class-wise precision, recall, and F1 score for real and fake classes, Matthews correlation coefficient (MCC), equal error rate (EER), PR-AUC for both classes, Brier score, and confusion matrices. The model output is denoted by $s \in [0, 1]$, where s is the sigmoid-bounded fake class decision score. Since no post hoc calibration was fitted in this study, s should be interpreted as a ranking and decision score rather than as a posterior probability.

The Brier score was computed as the mean squared error between s and the binary ground-truth label, with lower values indicating better score quality. The fake class PR-AUC was computed while using s as the ranking score. The real class PR-AUC was computed while using $1 - s$ as the real class score. The real class was treated as class 0, and the fake class was class 1. In the confusion matrices, true negatives (TNs) correspond to real samples correctly classified as real, false positives (FPs) correspond to real samples incorrectly classified as fake, false negatives (FNs) correspond to fake samples incorrectly classified as real, and true positives (TPs) correspond to fake samples correctly classified as fake.

Unless otherwise stated, the default decision threshold is $t = 0.5$. The high-threshold operating point $t^* = 0.98$ is reported as a frozen source-domain operating point that was fixed before DFDC testing and then applied unchanged to the test reports. The DFDC test set was not used for threshold selection, model selection, early stopping, optimizer selection, loss selection, focal-level selection, or any hyperparameter choice. No DFDC metric was used to resolve ties or near-ties. To avoid any ambiguity, the DFDC threshold sweep reported in Section 4.2 is presented only as a post hoc operating-point sensitivity analysis. Threshold-independent metrics, especially the ROC-AUC, EER, and PR-AUC, are treated as the primary evidence of score separability. Because the complete source-domain threshold sweep trajectory was not re-estimated in this revision, $t^* = 0.98$ was treated as a frozen source-domain operating point rather than as a threshold optimized or justified on the DFDC.

To provide a measure of finite test set uncertainty without retraining, we additionally report the stratified bootstrap 95% confidence intervals for the main threshold-dependent DFDC metrics. For each bootstrap replicate, real and fake samples were resampled with replacement within their respective classes, preserving the original class counts. The accuracy, balanced accuracy, macro F1 score, MCC, real recall, and fake recall were recomputed for each replicate, and the 2.5th and 97.5th percentiles are reported as the confidence interval.

These intervals quantify test set sampling uncertainty for the archived predictions; they do not replace repeated training with independent random seeds.

3.1.6. Model Size and Inference Time

Model complexity was measured from the verified checkpoints using the same pre-extracted feature inputs used during evaluation: video features (1, 14, 384) in size and audio features (1, 128, 768) in size at a batch size of 1. FLOPs were computed using the thop library, and the inference time was measured on the same NVIDIA RTX 6000 Ada Generation GPU by averaging 100 forward passes. As shown in Table 2, FocalNet introduced 0.435M additional trainable parameters relative to the reproduced 2D3MF baseline (+8.2%). Despite this increase, it used slightly fewer FLOPs (0.633 vs. 0.636 GFLOPs) and reduced the per-sample inference time from 3.04 ms to 2.57 ms. This speed improvement is attributed to replacing token-to-token softmax attention with focal modulation and cache-efficient depth-wise temporal convolutions.

Table 2. Model efficiency comparison measured from the verified checkpoints. FLOPs were computed for a batch size of 1 using video input (1, 14, 384) and audio input (1, 128, 768); inference time was averaged over 100 forward passes on an NVIDIA RTX 6000 Ada Generation GPU.

Metric	2D3MF Baseline	FocalNet	Difference
FLOPs (GFLOPs)	0.636	0.633	−0.003 (−0.5%)
Trainable parameters (M)	5.304	5.739	+0.435 (+8.2%)
Inference time (ms/sample)	3.04	2.57	15% faster
GPU	RTX 6000 Ada	RTX 6000 Ada	Same
DFDC AUC	0.5000	0.9324	+0.4324
DFDC BAcc at $t^* = 0.98$	0.5000	0.8030	+0.3030

3.2. Baseline: 2D3MF Reproduction

The 2D3MF architecture [8] was strictly followed in the reproduction of the baseline model. The baseline model performed cross-modal fusion via visual features extracted from MARLIN and audio features extracted from EAT, utilizing the cross-attention-based middle fusion process. In particular, the visual features were projected into a 384-to-128-dimensional space by applying a learnable projection operator. After that, two cross-attention mechanisms were applied, specifically (1) an audio-to-video ($A \rightarrow V$) cross-attention block, where the audio features serve as queries while the video features are used as keys and values; and (2) a video-to-audio ($V \rightarrow A$) cross-attention block, where the directions are reversed, i.e., the video features become queries while the audio features become keys and values. This bidirectional design helped the model understand the intermodal dependencies between both modalities.

The fused representation was then processed using 1-D CNN pooling layers to aggregate local temporal features. The output from the pooling layer was then globally average-pooled, concatenated, and processed by a binary classification layer using the BCE loss.

3.3. Proposed Method: Focal Modulation for Cross-Modal Fusion

The proposed framework modifies the middle-fusion stage of the reproduced 2D3MF architecture by replacing the softmax-based pairwise cross-attention module with a cross-modal focal modulation (CMFM) module. The purpose of this replacement is to reduce dependency on sharp token-to-token attention distributions and to encourage a smoother exchange of audio-visual contextual information under cross-dataset distribution shifts. The proposed module follows a first-aggregate, last-interact design, where multi-scale

temporal context is first extracted from one modality and is then used to modulate the other modality by element-wise interaction.

Let $\mathbf{X}_v \in \mathbb{R}^{T_v \times C_v}$ and $\mathbf{X}_a \in \mathbb{R}^{T_a \times C_a}$ denote the processed visual and audio feature sequences, respectively. In this work, $T_v = 14$, $C_v = 128$, $T_a = 128$, and $C_a = 768$ after the projection and positional encoding steps described previously. Given these two feature sequences, the proposed model learns the binary decision function $f_\theta(\mathbf{X}_v, \mathbf{X}_a)$.

$$s = f_\theta(\mathbf{X}_v, \mathbf{X}_a), \quad (1)$$

where s is the sigmoid-bounded fake class decision score for the input audio-visual sample and θ denotes the trainable parameters of the fusion and classification modules.

3.3.1. Motivation for Cross-Attention-Free Focal Fusion

In the replicated 2D3MF baseline, cross-modal interaction is modeled through cross-attention. Given a query sequence $\mathbf{Q}$, key sequence $\mathbf{K}$, and value sequence $\mathbf{V}$, conventional scaled dot-product attention can be written as follows;

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \quad (2)$$

where d is the feature dimensionality. Although this mechanism is effective for learning pairwise token relationships, the softmax operation may produce highly peaked attention weights. In cross-dataset evaluation, such peaked weights can make the model overly dependent on dataset-specific token correlations, which limits transferability to unseen data.

To address this issue, we replaced cross-attention with focal modulation [9]. Instead of computing explicit pairwise attention scores, the proposed CMFM block first aggregates contextual evidence from the source modality through multi-scale depth-wise temporal convolutions and then modulates the target modality through point-wise multiplication. This enables the model to capture both local temporal irregularities and broader cross-modal structural patterns without computing a token-to-token softmax attention matrix.

3.3.2. Cross-Modal Focal Modulation Block

Let $\mathbf{Q} \in \mathbb{R}^{T_q \times C_q}$ be the target sequence to be modulated, and let $\mathbf{K} \in \mathbb{R}^{T_k \times C_k}$ be the context sequence of the other modality. The CMFM block is denoted by $\mathcal{F}(\mathbf{Q}, \mathbf{K})$ and consists of three operations: context projection, multi-scale context aggregation with learnable focal-level gating, and element-wise modulation.

First, the context sequence is projected into the feature dimension of the target sequence:

$$\mathbf{Z}_c = \mathbf{K}\mathbf{W}_c + \mathbf{b}_c, \quad \mathbf{Z}_c \in \mathbb{R}^{T_k \times C_q}, \quad (3)$$

where $\mathbf{W}_c \in \mathbb{R}^{C_k \times C_q}$ and $\mathbf{b}_c$ are learnable parameters. The projected context is then passed through the focal levels L of one-dimensional depth-wise temporal convolutions:

$$\mathbf{C}_\ell = \text{DWConv}_{k_\ell}(\mathbf{Z}_c), \quad \ell = 1, 2, \dots, L, \quad (4)$$

where $\text{DWConv}_{k_\ell}(\cdot)$ denotes a temporal depth-wise convolution with a kernel size k_ℓ . In our implementation, $L = 3$, and $\{k_1, k_2, k_3\} = \{2, 4, 6\}$, enabling the model to aggregate context at multiple temporal scales. In addition to the local focal branches, a global context branch is computed by applying global average pooling along the temporal dimension:

$$\mathbf{c}_g = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathbf{Z}_c(t, :), \quad (5)$$

$$\mathbf{C}_g = \text{Broadcast}(\mathbf{c}_g), \quad \mathbf{C}_g \in \mathbb{R}^{T_k \times C_q}. \quad (6)$$

The local and global context branches are adaptively combined through a learnable gating function. The gate tensor is computed as follows:

$$\mathbf{G} = \sigma(\mathbf{Z}_c \mathbf{W}_g + \mathbf{b}_g), \quad \mathbf{G} \in \mathbb{R}^{T_k \times (L+1)C_q}, \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function. The tensor $\mathbf{G}$ is split along the channel dimension into $L + 1$ gates:

$$\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_L, \mathbf{G}_g], \quad (8)$$

where each gate has the same dimension as the projected context. The final multi-scale contextual modulator is obtained as follows:

$$\mathbf{M} = \sum_{\ell=1}^L \mathbf{G}_\ell \odot \mathbf{C}_\ell + \mathbf{G}_g \odot \mathbf{C}_g, \quad (9)$$

where $\odot$ denotes element-wise multiplication. Unlike cross-attention, this aggregation does not require a token-to-token similarity matrix or softmax-normalized pairwise attention weights.

Because the target and context modalities may have different temporal lengths, $\mathbf{M}$ is interpolated to the temporal length of the target sequence:

$$\bar{\mathbf{M}} = \mathcal{I}_{T_q}(\mathbf{M}), \quad \bar{\mathbf{M}} \in \mathbb{R}^{T_q \times C_q}, \quad (10)$$

where $\mathcal{I}_{T_q}(\cdot)$ denotes a one-dimensional temporal interpolation to a length T_q . The target sequence is then projected and modulated as follows:

$$\mathbf{U} = (\mathbf{Q} \mathbf{W}_q + \mathbf{b}_q) \odot \bar{\mathbf{M}}, \quad (11)$$

where $\mathbf{W}_q$ and $\mathbf{b}_q$ are learnable projection parameters. The output of the CMFM block is computed as follows:

$$\mathcal{F}(\mathbf{Q}, \mathbf{K}) = \text{LN}(\mathbf{U} \mathbf{W}_o + \mathbf{b}_o) + \rho(\mathbf{Q}), \quad (12)$$

where $\text{LN}(\cdot)$ denotes layer normalization, $\mathbf{W}_o$ and $\mathbf{b}_o$ are output projection parameters, and $\rho(\cdot)$ is a residual mapping. If the input and output dimensions match, then $\rho(\mathbf{Q}) = \mathbf{Q}$; otherwise, a linear projection is used to align the dimensions.

Table 3 summarizes the CMFM computation in pseudocode form. This first citation appears after Tables 1 and 2, preserving numerical order.

Table 3. Pseudocode summary of the cross-modal focal modulation (CMFM) block. The same block is used in both directions; audio-to-video modulation treats the audio sequence as context for the video sequence, while video-to-audio modulation reverses the target and context roles.

Step	Operation
1	Input target sequence $\mathbf{Q} \in \mathbb{R}^{T_q \times C_q}$, context sequence $\mathbf{K} \in \mathbb{R}^{T_k \times C_k}$, and focal kernels $\{k_1, \dots, k_L\}$.
2	Project the context sequence to the target channel dimension: $\mathbf{Z}_c \leftarrow \text{Linear}_{C_k \rightarrow C_q}(\mathbf{K})$.
3	For each focal level ℓ , compute a local context branch using depth-wise temporal convolution: $\mathbf{C}_\ell \leftarrow \text{DWConv}_{k_\ell}(\mathbf{Z}_c)$.
4	Compute a global context branch by temporal average pooling and broadcasting: $\mathbf{C}_g \leftarrow \text{Broadcast}(\text{AvgPool}_t(\mathbf{Z}_c))$.
5	Predict focal gates and split them into local and global gates: $[\mathbf{G}_1, \dots, \mathbf{G}_L, \mathbf{G}_g] \leftarrow \sigma(\text{Linear}(\mathbf{Z}_c))$.
6	Fuse local and global context branches: $\mathbf{M} \leftarrow \sum_{\ell=1}^L \mathbf{G}_\ell \odot \mathbf{C}_\ell + \mathbf{G}_g \odot \mathbf{C}_g$.
7	Interpolate the modulator to the target temporal length: $\bar{\mathbf{M}} \leftarrow \text{Interpolate}(\mathbf{M}, T_q)$.
8	Project and modulate the target sequence: $\mathbf{U} \leftarrow \text{Linear}(\mathbf{Q}) \odot \bar{\mathbf{M}}$.
9	Return the residual normalized output: $\mathcal{F}(\mathbf{Q}, \mathbf{K}) \leftarrow \text{LN}(\text{Linear}(\mathbf{U})) + \rho(\mathbf{Q})$.

3.3.3. Bidirectional Audio-Visual Modulation

To preserve the bidirectional interaction strategy of 2D3MF while replacing pairwise cross-attention, two CMFM blocks were used. In the audio-to-video branch, the audio representation provides contextual evidence to modulate the visual sequence:

$$\mathbf{Z}_v = \mathcal{F}_{a \rightarrow v}(\mathbf{X}_v, \mathbf{X}_a), \quad \mathbf{Z}_v \in \mathbb{R}^{T_v \times C_v}. \quad (13)$$

This branch allows the visual representation to be refined using speech-related context, which is important for detecting inconsistencies between facial motion and the audio signal.

Similarly, in the video-to-audio branch, the visual sequence provides context for modulating the audio representation:

$$\mathbf{Z}_a = \mathcal{F}_{v \rightarrow a}(\mathbf{X}_a, \mathbf{X}_v), \quad \mathbf{Z}_a \in \mathbb{R}^{T_a \times C_a}. \quad (14)$$

This branch enables the audio stream to be conditioned on facial dynamics and visual temporal cues. The two modulated outputs are therefore complementary; $\mathbf{Z}_v$ represents visual evidence conditioned with audio, while $\mathbf{Z}_a$ represents video-conditioned audio evidence.

3.3.4. Temporal Aggregation and Classification

After bidirectional focal modulation, each modality is processed by a separate one-dimensional convolutional aggregation module:

$$\mathbf{H}_v = \Psi_v(\mathbf{Z}_v), \quad \mathbf{H}_a = \Psi_a(\mathbf{Z}_a), \quad (15)$$

where $\Psi_v(\cdot)$ and $\Psi_a(\cdot)$ denote the visual and audio 1-D CNN grouping blocks, respectively. Global average pooling is then applied to obtain modality-level descriptors:

$$\mathbf{h}_v = \frac{1}{T_v} \sum_{t=1}^{T_v} \mathbf{H}_v(t, :), \quad (16)$$

$$\mathbf{h}_a = \frac{1}{T_a} \sum_{t=1}^{T_a} \mathbf{H}_a(t, :). \quad (17)$$

The final multimodal feature vector is obtained via concatenation:

$$\mathbf{h} = [\mathbf{h}_v \parallel \mathbf{h}_a], \quad (18)$$

where $\parallel$ denotes channel-wise concatenation. The classification head maps $\mathbf{h}$ to a scalar logit:

$$z = \mathbf{w}_c^\top \phi(\mathbf{W}_h \mathbf{h} + \mathbf{b}_h) + b_c, \quad (19)$$

where $\phi(\cdot)$ is a nonlinear activation function. The final fake class score is computed using the sigmoid function:

$$s = \sigma(z) = \frac{1}{1 + \exp(-z)}. \quad (20)$$

The predicted class is then defined as

$$\hat{y} = \begin{cases} 1, & s \geq t, \\ 0, & s < t, \end{cases} \quad (21)$$

where t is the decision threshold. During the default evaluation, $t = 0.5$ was used. The high-threshold point $t^* = 0.98$ is reported as a frozen source-domain operating point and is applied unchanged to the test reports. It was not selected or justified using DFDC metrics.

3.3.5. Training Objective and Implementation

The proposed model was trained using binary cross-entropy with logits. For a mini-batch of N samples, the optimization objective is

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(s_i) + (1 - y_i) \log(1 - s_i)], \quad (22)$$

where $y_i \in \{0, 1\}$ is the ground truth label and s_i is the sigmoid-bounded fake class decision score. Equation (22) is written in probability form for readability; in implementation, `BCEWithLogitsLoss` applies the same binary cross-entropy objective directly to logits using a numerically stable formulation, which is particularly useful under mixed-precision training.

The CMFM module was implemented as a drop-in replacement for the original attention block in the 2D3MF pipeline. The remaining components, including the 1-D CNN pooling layers, global pooling, classification head, optimizer, learning rate schedule, and training protocol, remained unchanged. This ensured that the comparison isolated the effect of replacing token-to-token softmax-based cross-attention with focal modulation. In our implementation, the CMFM block used $L = 3$ focal levels with kernel sizes $\{2, 4, 6\}$. Under the verified feature-level implementation, FocalNet contained 5.739M trainable parameters, compared with 5.304M for the reproduced 2D3MF baseline, corresponding to an increase of 0.435M parameters (+8.2%), while using 0.633 GFLOPs compared with 0.636 GFLOPs for the baseline.

3.4. Additional Diagnostic Ablation Variants

In addition to the optimizer/loss ablation, three targeted diagnostic variants were trained to evaluate whether simpler gating or imbalance-handling modifications could improve the final model. These variants were not used for threshold selection or model selection on the DFDC; the DFDC remained an unseen target-domain test set.

Experiment 1: FocalNet-SG. This variant replaced the original focal-level gating formulation with independent sigmoid gates for the focal branches. It kept the original class distribution, Adam optimizer, and BCE loss. The purpose was to test whether a fully independent sigmoid-gated focal module improved transfer compared with the original FocalNet gating design.

Experiment 2: FocalNet-SG-Bal. This variant used the independent sigmoid-gated focal module and added a weighted training sampler based on the inverse class frequency in the FakeAVCeleb training split. The goal was to test whether direct sampling-based class balancing could improve the minority real-class recall without changing the loss function.

Experiment 3: FocalNet-SG-Bal-MRG. This variant extended Experiment 2 with a lightweight residual modality reliability gate after modality-level pooling. The gate used the concatenated audio-visual descriptor to softly rescale the audio and video descriptors before final classification. The residual form prevented either modality from being completely suppressed and tested whether reliability-aware fusion could recover the fake class performance lost by balanced sampling.

These three experiments were intentionally treated as diagnostic ablations rather than replacements for the final model. Their role was to clarify whether the performance of the original FocalNet could be replicated by naive independent gating, class balancing, or a simple modality reliability gate.

3.5. Cross-Modal Explainability Analysis

3.5.1. Method

To inspect the model decision process, we used Integrated Gradients (IG) [43] on the pre-extracted MARLIN and EAT feature representations. For feature dimension i , IG is defined as

$$\text{IG}_i(\mathbf{x}) = (x_i - x'_i) \int_0^1 \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} d\alpha, \quad (23)$$

where $\mathbf{x}$ is the input feature vector, $\mathbf{x}'$ is the baseline input, $F(\cdot)$ is the scalar model output used for attribution, x_i is the i th feature dimension, and α is the interpolation coefficient along the straight-line path from the baseline to the input. We used a zero-vector baseline and approximated the integral with a 50-step Riemann sum.

IG was computed with respect to the pre-sigmoid fake class logit rather than the saturated sigmoid score. This choice avoided artificially suppressing gradients for samples with scores extremely close to 0 or 1. The resulting attributions are then summarized only as feature-level diagnostics. We did not use the IG magnitude as a calibrated measure of causal importance, and we did not compare the raw attribution magnitude across samples as a quantitative performance metric.

For temporal interpretability, the ℓ_2 -norm of the attribution scores was calculated at each temporal position, yielding one importance value per position for each modality. This produced attribution sequences of a length $T = 14$ for the visual stream and 512 frames for the audio stream; the audio sequence was pooled into 20 temporal windows only for descriptive plotting and summary. The attribution analysis was performed on a small post hoc subset of the DFDC test set consisting of 30 real and 30 fake samples using the final FocalNet checkpoint. This subset was used only for qualitative and descriptive inspection.

3.5.2. Scope and Constraints

The revised analysis does not assign contiguous MARLIN embedding dimensions to anatomical variables. MARLIN uses region-guided masking during pretraining, but this does not imply that fixed contiguous embedding channels correspond to anatomically ordered facial regions. Therefore, anatomical attribution claims were removed, and the IG results are interpreted only as feature-channel, temporal, and modality-level attributions. Because IG was applied to pre-extracted features rather than to raw pixels or spatial tokens, the analysis should not be interpreted as pixel-level image mapping, image-space visualization, or evidence of specific anatomical facial phenomena. A valid image-space attribution analysis would require backpropagation through the MARLIN encoder to known tokens, patches, or pixels, which is outside the scope of the present feature-level study.

4. Results

4.1. Imbalance-Aware Cross-Dataset Evaluation

The comprehensive evaluation is reported in Table 4. Compared with the previous summary table, the revised table includes the balanced accuracy, MCC, EER where available, and class-wise F1 scores. This is important because the FakeAVCeleb source splits were highly imbalanced, and the DFDC target test set was also fake-dominant. Under such conditions, high accuracy can be produced by a classifier that simply predicts the majority class.

The reproduced 2D3MF baseline illustrates this problem clearly. On the DFDC, the baseline achieved an apparent accuracy of 0.8313 because 4323 of the 5200 DFDC samples were fake. However, the confusion matrix in Table 5 shows that the baseline predicted every DFDC sample as fake. Consequently, it obtained a real-class recall zero, real-class F1 score of zero, balanced accuracy of 0.5000, MCC of zero, and macro F1 score

of 0.4540. This corrected result replaces the earlier recall summary and illustrates why accuracy and ROC-AUC alone are insufficient for this task.

In the reported seed-42 evaluation, replacing cross-attention with cross-modal focal modulation produced stronger cross-dataset score separability than the reproduced baseline. At the default threshold $t = 0.5$, FocalNet obtained a DFDC AUC of 0.9324, macro F1 score of 0.7553, balanced accuracy of 0.7003, and MCC of 0.5857. The model correctly identified 4310 of 4323 fake samples while also recovering 354 of 877 real samples. At the frozen high-threshold operating point $t^* = 0.98$, the real-class recall increased from 0.4036 to 0.6306, and the macro F1 score was 0.8358. This operating point also yielded a balanced accuracy of 0.8030 and MCC of 0.6821 while maintaining a high fake class recall of 0.9755. The corresponding stratified bootstrap intervals in Table 6 show that these thresholded improvements remained separated from the all-fake baseline under finite test-set resampling while still not replacing multi-seed training. The high-threshold point is reported as a pre-specified operating point, whereas the DFDC threshold sweep in Section 4.2 is only a sensitivity analysis.

Table 4. Comprehensive results under the FakeAVCeleb $\rightarrow$ DFDC protocol. BAcc denotes balanced accuracy, and MCC denotes the Matthews correlation coefficient. PR-AUC values are reported separately in the threshold-sensitivity analysis.

Model	Dataset	Threshold	ACC	AUC	EER	BAcc	MCC	R_{real}	R_{fake}	$F1_{real}$	$F1_{fake}$	Macro F1	Notes
2D3MF baseline	FakeAVCeleb	$t = 0.5$	0.9780	0.8135	—	0.5295	0.1849	0.0600	0.9991	0.1091	0.9889	0.5490	In-domain baseline
2D3MF baseline	DFDC	$t = 0.5$	0.8313	0.5000	0.5000	0.5000	0.0000	0.0000	1.0000	0.0000	0.9079	0.4540	All-fake prediction
FocalNet	FakeAVCeleb	$t = 0.5$	0.9780	0.9385	0.1207	0.5100	0.1399	0.0200	1.0000	0.0392	0.9889	0.5140	In-domain proposed
FocalNet	DFDC	$t = 0.5$	0.8969	0.9324	0.1494	0.7003	0.5857	0.4036	0.9970	0.5691	0.9415	0.7553	Default threshold
FocalNet	DFDC	$t^* = 0.98$	0.9173	0.9324	0.1494	0.8030	0.6821	0.6306	0.9755	0.7201	0.9515	0.8358	Frozen high-threshold point

Note: “—” indicates that the EER value was unavailable from the archived baseline evaluation output.

Table 5. DFDC confusion matrices for the reproduced baseline and the proposed FocalNet. TN = real predicted as real; FP = real predicted as fake; FN = fake predicted as real; TP = fake predicted as fake.

Model	Threshold	Dataset	TN	FP	FN	TP
2D3MF baseline	$t = 0.5$	DFDC	0	877	0	4323
FocalNet	$t = 0.5$	DFDC	354	523	13	4310
FocalNet	$t^* = 0.98$	DFDC	553	324	106	4217

Table 6. Stratified bootstrap 95% confidence intervals for the main DFDC threshold-dependent metrics. Intervals were computed by resampling the archived thresholded predictions within the real and fake classes; they quantify finite test-set sampling uncertainty and do not measure training-seed variability.

Metric	2D3MF Baseline, $t = 0.5$	FocalNet, $t = 0.5$	FocalNet, $t^* = 0.98$
ACC	0.8313 [0.8313, 0.8313]	0.8969 [0.8913, 0.9025]	0.9173 [0.9106, 0.9238]
BAcc	0.5000 [0.5000, 0.5000]	0.7003 [0.6841, 0.7166]	0.8030 [0.7868, 0.8191]
MCC	0.0000 [0.0000, 0.0000]	0.5857 [0.5579, 0.6129]	0.6821 [0.6544, 0.7093]
Macro F1	0.4540 [0.4540, 0.4540]	0.7553 [0.7370, 0.7728]	0.8358 [0.8211, 0.8500]
R_{real}	0.0000 [0.0000, 0.0000]	0.4036 [0.3717, 0.4356]	0.6306 [0.5986, 0.6625]
R_{fake}	1.0000 [1.0000, 1.0000]	0.9970 [0.9951, 0.9986]	0.9755 [0.9709, 0.9801]

4.2. Threshold Sensitivity and Raw Score Analysis

To document the operating-point sensitivity under severe class imbalance, we performed a post hoc threshold sweep on the FocalNet DFDC scores from $t = 0.40$ to $t = 0.99$ in steps of 0.01. This sweep was not a threshold selection procedure and was not used to choose $t^* = 0.98$. Instead, it was included to make the dependence of thresholded

metrics on the decision boundary transparent. All architecture, optimizer, loss, training, and threshold decisions were finalized without using DFDC labels.

The threshold-independent metrics from the full DFDC score distribution were ROC-AUC = 0.9324, EER = 0.1494, real class PR-AUC = 0.8205, and fake class PR-AUC = 0.9830. These metrics are the primary evidence that the scores retained real and fake separability under the FakeAVCeleb-to-DFDC shift. Table 7 reports the selected operating points only to illustrate the trade-off between real recall and fake recall. In particular, $t = 0.99$ yielded a slightly higher DFDC macro F1 score and balanced accuracy than $t^* = 0.98$, while $t^* = 0.98$ gave a marginally higher MCC. This confirms the reviewer's observation that thresholded conclusions depend on the chosen operating point. For this reason, the revised manuscript avoids presenting the DFDC MCC as the basis for choosing or justifying the main threshold.

Table 7. Post hoc DFDC operating point sensitivity for FocalNet. PR-AUC values were threshold-independent and equaled 0.8205 for the real class and 0.9830 for the fake class. This table was not used for threshold selection.

Threshold	TN/FP/FN/TP	ACC	BACC	MCC	R_{real}	R_{fake}	$F1_{\text{real}}$	$F1_{\text{fake}}$	Macro F1
0.50	354/523/13/4310	0.8969	0.7003	0.5857	0.4036	0.9970	0.5691	0.9415	0.7553
0.80	415/462/25/4298	0.9063	0.7337	0.6289	0.4732	0.9942	0.6302	0.9464	0.7883
0.90	450/427/35/4288	0.9112	0.7525	0.6503	0.5131	0.9919	0.6608	0.9488	0.8048
0.95	488/389/67/4256	0.9123	0.7705	0.6560	0.5564	0.9845	0.6816	0.9492	0.8154
0.98	553/324/106/4217	0.9173	0.8030	0.6821	0.6306	0.9755	0.7201	0.9515	0.8358
0.99	606/271/175/4148	0.9142	0.8253	0.6818	0.6910	0.9595	0.7310	0.9490	0.8400

4.3. Score Quality and Calibration Scope

Raw model outputs were used in this study as decision scores rather than posterior probabilities. For this reason, the revised manuscript reports the Brier score as a simple score quality diagnostic and interprets it together with confusion matrices. The expected calibration error (ECE) was not used as a main quantitative result in the revised manuscript because the previously reported baseline ECE under degenerate all-fake predictions was sensitive to implementation and binning details. Without per-bin counts, average confidence values, empirical accuracies, and reliability diagrams, such an ECE value could be misleading.

Table 8 therefore reports the Brier scores only. The reproduced 2D3MF baseline obtained a DFDC Brier score of 0.1687 and predicted every DFDC sample as fake, which is consistent with poor target-domain score behavior despite high apparent accuracy. FocalNet reduced the DFDC Brier score to 0.0932, indicating better raw score quality under the evaluated cross-dataset setting. However, this should not be interpreted as evidence of posterior probability calibration. A dedicated calibration analysis, including an explicit ECE equation, bin assignment rules for scores at zero or one, per-bin counts, reliability diagrams, and possible calibration methods such as temperature scaling, is left for future work.

Table 8. Score quality metrics computed from raw per-sample prediction scores. The Brier score was interpreted together with the confusion matrices because the scores were not treated as posterior probabilities.

Model	Dataset	Brier Score	Interpretation
2D3MF baseline	DFDC	0.1687	All-fake prediction pattern
2D3MF baseline	FakeAVCeleb	0.0205	In-domain score quality
FocalNet	DFDC	0.0932	Better DFDC score quality than baseline
FocalNet	FakeAVCeleb	0.0211	In-domain score quality

4.4. In-Domain and Cross-Dataset Behavior

On FakeAVCeleb, both the baseline and FocalNet obtained high accuracy because the test split contained only 50 real samples and 2175 fake samples. Therefore, the in-domain results should be interpreted with class imbalance in mind. The reproduced 2D3MF baseline obtained an in-domain AUC of 0.8135, while FocalNet improved the in-domain AUC to 0.9385. However, the cross-dataset results are more informative. The reproduced 2D3MF baseline failed to transfer to the DFDC and collapsed into the majority fake class. FocalNet instead produced higher score separation between real and fake samples on the unseen DFDC distribution in the reported seed-42 run, as shown by its AUC of 0.9324 and improved MCC.

The threshold-adjusted result shows that FocalNet's prediction scores can be adjusted to better balance the minority real class and the majority fake class. At $t = 0.5$, the model was conservative with respect to the real class and obtained a real recall of 0.4036. At $t^* = 0.98$, the real recall increased to 0.6306, while the fake recall remained high at 0.9755. This trade-off is more transparent when evaluated using confusion matrices, balanced accuracy, MCC, and the class-wise F1 score rather than accuracy alone.

4.5. Feature-Level Explainability Analysis

To interpret the FocalNet decision-making process, we used Integrated Gradients on the pre-extracted MARLIN and EAT feature sequences. The revised analysis was intentionally narrower than the previous version; it was limited to temporal and modality-level feature attribution and did not report anatomical region attribution, image-space overlays, or localized facial artifact claims.

4.5.1. Visual Feature Temporal Attribution

For the analyzed DFDC subset, fake samples showed relatively stronger attribution in the earliest visual feature segments, especially the first two MARLIN temporal positions. This indicates that those temporal feature segments contributed more strongly to the pre-sigmoid fake class logit for these examples. Because the analysis was performed on pre-extracted MARLIN features, this result should be interpreted only as temporal feature attribution. It does not identify the responsible pixels, facial parts, or local image artifacts.

Real samples showed a more distributed visual feature attribution pattern across the sampled temporal sequence. This suggests that, in the inspected examples, the model used broader visual temporal evidence when assigning lower fake class scores. The result is descriptive and post hoc and not causal evidence of temporal reasoning.

4.5.2. Audio Feature Temporal Attribution

The audio attribution analysis suggests that real samples had larger attribution scores across several audio windows than fake samples in the inspected subset. This indicates that the audio feature stream contributed more visibly for genuine samples in this post hoc analysis, whereas fake samples were associated with relatively greater visual feature attribution. Because the model outputs are uncalibrated scores, and the analysis was based on only 30 real and 30 fake examples, the results should not be interpreted as a calibrated modality-reliability estimate or as a statistically general explanation of all DFDC videos.

4.5.3. Interpretation Scope

The revised explainability section deliberately avoided assigning MARLIN channels to anatomical facial regions. It also removed claims about specific anatomical or image-space phenomena that cannot be supported by feature-level IG. The retained conclusion is limited to the following: in the analyzed examples, FocalNet's fake class logit was associated

with early visual feature segments and lower audio feature attribution, while real samples showed more distributed temporal feature attribution and greater audio contribution. A valid image-space explanation would require attribution through the MARLIN encoder to known tokens, patches, or pixels and would also require reporting the completeness error, normalization procedures, and uncertainty across samples. Those analyses are outside the scope of the present feature-level study.

4.6. Ablation Study: Optimizer and Loss Function

To examine whether the gain of FocalNet was due merely to optimizer or loss function choices, we held the FocalNet architecture fixed and changed only the optimizer and loss. The four settings were Adam + BCE, SGDM + BCE, Adam + focal loss, and SGDM + focal loss. All variants were trained on FakeAVCeleb and evaluated as zero-shot on the DFDC. Table 9 reports the AUC, EER, class recalls, balanced accuracy, macro F1 score, MCC, and class-wise F1 scores.

Table 9. Optimizer/loss ablation on the DFDC with the FocalNet architecture fixed. Counts were derived from the verified DFDC confusion matrices.

Exp.	Setting	AUC	EER	R_{real}	R_{fake}	BAcc	Macro F1	MCC	Real F1	Fake F1	FAV AUC
A	Adam + BCE	0.9324	0.1494	0.4036	0.9970	0.7003	0.7553	0.5857	0.5691	0.9415	0.9385
B	SGDM + BCE	0.7480	0.3125	0.0182	0.9993	0.5088	0.4724	0.1089	0.0357	0.9091	0.7658
C	Adam + Focal Loss	0.8284	0.2487	0.4983	0.8980	0.6981	0.6980	0.3961	0.4980	0.8981	0.8636
D	SGDM + Focal Loss	0.8188	0.2532	0.5656	0.8670	0.7163	0.6981	0.4005	0.5092	0.8869	0.8188
Ref.	2D3MF baseline	0.5000	0.5000	0.0000	1.0000	0.5000	0.4540	0.0000	0.0000	0.9079	0.8135

Adam + BCE provided the best cross-dataset ranking performance among the reported single-seed settings, with a DFDC AUC of 0.9324. Replacing Adam with SGDM reduced the AUC to 0.7480 and almost eliminated real class recovery ($R_{real} = 0.0182$), although the fake recall remained nearly perfect. Focal loss improved the real/fake recall balance; the Adam + focal loss reached $R_{real} = 0.4983$, and the SGDM + focal loss reached $R_{real} = 0.5656$. However, both focal loss settings reduced the AUC compared with Adam + BCE and also reduced the fake class recall. These results indicate that the focal loss and SGDM primarily modified the decision trade-off under imbalance, whereas the best cross-dataset separability among the reported single-seed settings was obtained by the focal modulation architecture trained with Adam and BCE.

4.7. Additional Architecture and Imbalance Ablations

Table 10 reports the three additional diagnostic experiments. These experiments are useful because they tested whether the main result can be improved by a simpler gate, direct class balancing, or a modality reliability gate. The answer is negative: none of the three variants surpassed the original FocalNet or the threshold-adjusted FocalNet result.

The independent sigmoid-gated variant, FocalNet-SG, performed poorly on the DFDC dataset. Its AUC decreased to 0.4754, and its MCC became negative (-0.0795), indicating that naive independent sigmoid gating damaged cross-dataset score separability. This result shows that the focal-level gating formulation is not a trivial implementation detail.

The balanced-sampler variant, FocalNet-SG-Bal, increased the real class recall to 0.8951 at $t = 0.5$, but it overcorrected the decision boundary and reduced the fake recall to 0.1906. Consequently, the macro F1 score and MCC remained low. At $t = 0.98$, the same model became even more real-biased, with a real recall of 0.9304 but a fake recall of only 0.1367. These results indicate that naive sampling-based balancing improved the minority sensitivity in this run but substantially weakened fake detection under cross-dataset shifts.

Table 10. Additional architecture and imbalance ablations on DFDC. All variants were trained on FakeAVCeleb and evaluated as zero-shot on DFDC. TN, FP, FN, and TP follow the conventions defined in Section 3.1.

Variant	t	AUC	BAcc	Macro F1	MCC	R_{real}	R_{fake}	Real F1	Fake F1	TN	FP	FN	TP
Original FocalNet	0.50	0.9324	0.7003	0.7553	0.5857	0.4036	0.9970	0.5691	0.9415	354	523	13	4310
Original FocalNet	0.98	0.9324	0.8030	0.8358	0.6821	0.6306	0.9755	0.7201	0.9515	553	324	106	4217
FocalNet-SG	0.50	0.4754	0.4553	0.4578	−0.0795	0.1562	0.7543	0.1320	0.7835	137	740	1062	3261
FocalNet-SG-Bal	0.50	0.5977	0.5429	0.3094	0.0842	0.8951	0.1906	0.3042	0.3146	785	92	3499	824
FocalNet-SG-Bal	0.98	0.5977	0.5336	0.2692	0.0759	0.9304	0.1367	0.3008	0.2376	816	61	3732	591
FocalNet-SG-Bal-MRG	0.50	0.6154	0.5783	0.4654	0.1181	0.6887	0.4680	0.3195	0.6113	604	273	2300	2023
FocalNet-SG-Bal-MRG	0.98	0.6154	0.5712	0.4100	0.1138	0.7925	0.3500	0.3172	0.5028	695	182	2810	1513

The modality reliability gate in FocalNet-SG-Bal-MRG partially corrected this imbalance. At $t = 0.5$, the fake recall improved from 0.1906 to 0.4680, and the MCC improved from 0.0842 to 0.1181 compared with FocalNet-SG-Bal. However, the variant remained far below the original FocalNet in terms of AUC, macro F1 score, and MCC. Increasing the threshold to $t = 0.98$ improved the real recall to 0.7925 but reduced the fake recall to 0.3500. Therefore, the original FocalNet trained with Adam and BCE remained the best reported single-seed configuration, while these additional experiments provided diagnostic evidence about architecture sensitivity and the risks of naive imbalance correction. These diagnostic variants are complemented by the focal-scale ablation reported next; the remaining global branch and direction-specific ablations are left for future work.

4.8. Component-Level Focal Scale Ablation

To examine the effect of reducing the local focal hierarchy, we trained controlled variants in which the number of local focal convolution branches was changed while the optimizer, loss, seed, training data, and evaluation protocol were held fixed. The global context branch was retained in all three variants; only the local focal branches were changed. The $L = 1$ variant used kernel $\{2\}$, the $L = 2$ variant used kernels $\{2, 4\}$, and the proposed $L = 3$ setting used kernels $\{2, 4, 6\}$. Table 11 reports the DFDC results at the default threshold $t = 0.5$.

Table 11. Component-level focal-scale ablation on the DFDC at the default threshold $t = 0.5$. All variants were trained on FakeAVCeleb with Adam + BCE and seed 42. The table is a diagnostic single-seed, fixed-threshold comparison rather than a complete architectural conclusion.

Variant	Kernels	AUC	ACC	BAcc	MCC	R_{real}	R_{fake}	Pattern at $t = 0.5$
$L = 1$	$\{2\}$	0.5988	0.169	0.5000	0.0000	1.000	0.000	All-real prediction
$L = 2$	$\{2, 4\}$	0.5724	0.169	0.5000	0.0000	1.000	0.000	All-real prediction
$L = 3$ (proposed)	$\{2, 4, 6\}$	0.9324	0.897	0.7003	0.5857	0.404	0.997	Better default-threshold balance

At the default threshold, the $L = 1$ and $L = 2$ variants predicted all DFDC samples as real. This produced a balanced accuracy of 0.5000 and a fake recall of zero, but it should not be interpreted as evidence that the outputs are completely non-separable. Their AUC values, 0.5988 and 0.5724, indicate limited but non-zero score separability. The main conclusion is therefore softened: under the present single-seed, fixed-threshold setting, reducing the number of local focal levels substantially weakened target-domain score separation and produced a poorly aligned operating point. A stronger architectural claim would require validation-selected thresholds, in-domain test reporting, and repeated multi-seed training for each ablation.

4.9. Protocol Context Relative to Prior DFDC-Reported Methods

Table 12 places the proposed result in the context of representative audio-visual methods that report the DFDC AUC. The table is not intended as a direct numerical ranking. The published DFDC results may differ in training datasets, DFDC split or subset, preprocessing, availability of audio, single-person filtering, and whether predictions were aggregated at the frame, segment, or video level. The present work used a filtered 5200-sample single-person DFDC subset and evaluated models trained only on FakeAVCeleb. Therefore, the reported AUC should be interpreted as contextual evidence under the stated protocol rather than as a claim of superiority over methods evaluated on different DFDC subsets.

Table 12. Protocol context for representative audio-visual deepfake detection methods reporting the DFDC AUC. The results are not directly comparable because the DFDC subsets, preprocessing, filtering, audio availability, and aggregation levels differ across studies.

Method	Year	Reported DFDC AUC	Training Data	DFDC Use/Subset	Protocol Note
Emotions Don't Lie [32]	2020	0.844	DFDC	DFDC used in training/testing	In-dataset AV result
AVoiD-DF [37]	2023	0.948	FakeAVCeleb + DefakeAVMiT + DFDC	DFDC used in training/testing	Dataset mixture differs
2D3MF [8]	2024	0.978	Not fully specified in cited report	Reported subset not matched here	Prior protocol differs
SADD [42]	2024	0.9669	FakeAVCeleb + DFDC	DFDC used in training/testing	In-/mixed-dataset AV result
AVT ² -DWF [38]	2024	0.8920	FakeAVCeleb + DeepfakeTIMIT + DFDC	DFDC used in training/testing	Dataset mixture differs
Baseline (2D3MF, reproduced)	2025	0.500	FakeAVCeleb only	Filtered 5200 single-person DFDC	Cross-dataset * AV result
FocalNet (ours)	2025	0.932	FakeAVCeleb only	Filtered 5200 single-person DFDC	Cross-dataset * AV result

* Cross-dataset denotes training without DFDC samples and testing on the filtered single-person DFDC subset used in this work.

The comparison remains useful because it highlights protocol differences. Several methods use the DFDC dataset during training or train on multiple datasets, whereas our reproduced baseline and FocalNet were trained only on FakeAVCeleb and evaluated as zero-shot on the filtered DFDC subset. Under this stricter source-to-target setting, the reproduced 2D3MF baseline collapsed to an all-fake prediction pattern, while FocalNet retained stronger score separability. However, because evaluation sets are not identical, the revised manuscript avoids describing the proposed result as directly numerically competitive with all prior DFDC-reported methods.

5. Discussion

5.1. Impact of Focal Modulation Under Distribution Shift

The revised analysis shows that the main difficulty is not only cross-dataset generalization but cross-dataset generalization under severe class imbalance. The reproduced 2D3MF baseline obtained a high apparent accuracy on the DFDC because the target test set was fake-dominant, yet its confusion matrix revealed complete failure on the minority real class. This is a stronger and more transparent finding than the original AUC- and accuracy-only summary: the baseline did not merely perform weakly; it collapsed to a majority-class prediction pattern.

FocalNet changed this behavior in the reported seed-42 evaluation. At the default threshold, it correctly identified nearly all fake DFDC samples while also recovering a non-trivial portion of real samples, improving the balanced accuracy and MCC relative to the baseline. At the frozen high-threshold operating point, the real class recall improved further while maintaining a high fake recall. This suggests that focal modulation is associated with more useful score separation under distribution shifts than the reproduced softmax-based cross-attention baseline, even though threshold choice remains important in imbalanced settings.

The additional diagnostic ablations clarify that this improvement was not obtained by any arbitrary focal-style modification. Independent sigmoid focal gating degraded the DFDC AUC below 0.5 and produced a negative MCC. Balanced sampling strongly

increased the real recall but collapsed the fake recall, while the residual modality reliability gate only partially recovered this loss. These results indicate that the reported single-seed cross-dataset transfer depends on the original focal modulation fusion and operating point choice rather than naive gating or sampling modifications.

Two aspects of the proposed architecture may explain this behavior. First, replacing pairwise cross-attention with contextual modulation reduces dependence on sharply normalized token relationships that may overfit to the source dataset. Second, the multi-scale temporal convolutions in the focal modulation block introduce a local-to-global temporal bias, enabling the model to aggregate short-term artifacts and broader audio-visual context before cross-modal interaction. Together, these properties may support more transferable audio-visual representations.

5.2. Class Imbalance, Score Quality, and Metric Transparency

The exact split counts confirm that FakeAVCeleb was highly imbalanced, with only 500 real samples compared with 21,044 fake samples after filtering. The test split contained only 50 real samples. Therefore, in-domain accuracy is not an adequate measure of performance. The DFDC test set was less extreme but remained fake-dominant, with 877 real and 4323 fake samples. These distributions explain why an all-fake classifier can obtain 83.13% accuracy on the DFDC subset while failing completely on real samples.

For this reason, the revised evaluation emphasized the balanced accuracy, MCC, EER, class-wise F1 score, and confusion matrices. The threshold-adjusted FocalNet result illustrates the value of this approach: the accuracy improvement was modest, but the real-class recall, balanced accuracy, MCC, and macro F1 score all improved substantially. In forensic settings, this distinction is important because a detector that incorrectly flags all real videos as fake is not practically useful, even if it obtains high accuracy on a fake-heavy benchmark.

The optimizer/loss and sampling ablations further show that imbalance-aware interventions can improve minority-class recall, but this does not necessarily translate into better cross-dataset ranking or score quality. Adam + focal loss and SGDM + focal loss improved the real recall relative to SGDM + BCE, but both reduced the AUC compared with Adam + BCE. Similarly, the balanced-sampler variants recovered many real samples but misclassified many fake samples as real. This indicates that class-sensitive objectives and sampling strategies should be tuned carefully; they may improve a single operating point while reducing overall score separability. The Brier score and confusion matrices are used as transparent score-quality diagnostics, while the model outputs are treated as uncalibrated classification scores rather than deployment-ready posterior probabilities.

5.3. Explainability Insights and Their Implications

The Integrated Gradients analysis provided a feature-level diagnostic of how FocalNet responds to audio and visual representations. In the analyzed subset, fake samples showed stronger attribution in early visual feature segments, while real samples showed more distributed temporal attribution and greater audio feature contribution. These patterns are useful for descriptive interpretation, but they should not be read as establishing pixel-level image mapping or anatomical-region evidence.

The revised manuscript therefore avoids interpreting contiguous MARLIN channels as anatomical facial regions. Because IG was applied to pre-extracted features rather than to raw frames or spatial tokens, the explanation results should be viewed as temporal and modality-level feature attribution only and not as causal evidence about specific facial structures or local image regions.

5.4. Limitations

Several limitations remain and are made explicit in the revised manuscript. First, the cross-dataset evaluation used one source dataset, FakeAVCeleb, and one unseen target subset, the DFDC. Testing on additional target datasets such as FaceForensics++, Celeb-DF-v2, KoDF, and DeeperForensics would provide a stronger estimate of generalization across manipulation methods, compression pipelines, identities, and acquisition conditions.

Second, the reported training results were based on the archived seed-42 runs. The added stratified bootstrap confidence intervals quantified uncertainty due to finite DFDC test set sampling for the archived predictions, but they did not measure optimization variability. The optimizer/loss ablations and focal-scale ablations therefore provided deterministic single-run evidence rather than statistically averaged estimates. Multi-seed training with repeated validation-based threshold selection is needed before making claims about statistical robustness or optimization stability.

Third, thresholded metrics are sensitive to the operating point. The frozen high-threshold point $t^* = 0.98$ was reported as an operating point fixed before DFDC testing, while the DFDC threshold sweep was used only for sensitivity analysis. Because different class priors or deployment requirements may favor different real and fake recall trade-offs, the ROC-AUC, EER, PR-AUC, confusion matrices, and threshold sweeps should be interpreted together.

Fourth, calibration was not fully evaluated. The revised manuscript removed ECE from the main comparison because the previous baseline value under a degenerate all-fake prediction pattern could not be interpreted without bin-level evidence. The reported Brier scores should therefore be viewed as score quality diagnostics and not as a complete calibration study.

Fifth, the explainability analysis was feature-level and post hoc. Since IG was applied to pre-extracted MARLIN and EAT features, the analysis did not establish anatomical region attribution, pixel-level image mapping, or causal reliance on specific facial structures. Finally, the focal-scale ablation was diagnostic; although the global context branch was retained across the $L = 1$, $L = 2$, and $L = 3$ variants, validation-selected thresholds, in-domain test results, and multi-seed runs are still required for a complete architectural conclusion.

5.5. Future Directions

Future work should extend the evaluation to additional unseen target datasets and manipulation types, including FaceForensics++, Celeb-DF-v2, KoDF, DeeperForensics, and in-the-wild social media videos. Multi-seed statistical testing should be performed, with the full validation-based threshold selection procedure repeated independently for each seed. Additional architecture-level ablations should examine removal of the global context branch, unidirectional versus bidirectional modulation, alternative focal kernels, and stronger reliability gates. Future deployment-oriented studies should analyze calibration, threshold transfer under changing class priors, robustness to compression and audio degradation, and throughput on resource-limited hardware. Finally, explainability should be extended from feature-level attribution to valid image-space attribution through known MARLIN tokens, patches, or pixels before any anatomical or image localization claims are made.

6. Conclusions

This study investigated cross-modal focal modulation for audio-visual deepfake detection under a strict FakeAVCeleb $\rightarrow$ DFDC protocol. The main finding is that replac-

ing softmax-based middle fusion with multi-scale focal modulation improved zero-shot score separation on the unseen DFDC subset while preserving practical computational efficiency. The reproduced 2D3MF baseline collapsed to a single-class prediction pattern under this protocol, whereas FocalNet produced more informative real/fake separation and more balanced thresholded behavior when the frozen source-domain operating point was applied.

The study also highlights several broader lessons for audio-visual deepfake evaluation. Accuracy alone is insufficient under strong class imbalance, thresholded results are sensitive to the chosen operating point, and feature-level explainability should not be over-interpreted as spatial or anatomical localization. For this reason, the revised evaluation emphasizes the class-wise recall, balanced accuracy, MCC, EER, PR-AUC, Brier score, confusion matrices, threshold sensitivity analysis, and stratified bootstrap confidence intervals for the main thresholded DFDC metrics.

Future work should test FocalNet on additional unseen datasets, including FaceForensics++, Celeb-DF-v2, KoDF, DeeperForensics, and in-the-wild social media videos. Multi-seed training is also needed to quantify optimization stability, and calibration-aware threshold transfer should be studied before deployment in settings with different class priors or operating costs. Finally, explainability should be extended beyond pre-extracted feature attribution toward pixel-level or token-level attribution through the visual encoder, enabling spatially grounded forensic interpretation.

Author Contributions: Conceptualization, S.M.B.D. and T.M.K.; methodology, S.M.B.D. and T.M.K.; software, S.M.B.D. and T.M.K.; validation, S.M.B.D. and T.M.K.; formal analysis, S.M.B.D. and T.M.K.; investigation, S.M.B.D. and T.M.K.; resources, S.M.B.D. and T.M.K.; data curation, S.M.B.D. and T.M.K.; writing—original draft preparation, S.M.B.D. and T.M.K.; writing—review and editing, Q.E.U.H.; visualization, Q.E.U.H.; supervision, Q.E.U.H.; project administration, Q.E.U.H.; funding acquisition, T.M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Naif Arab University for Security Sciences, grant number NAUSS-26-R05. The APC was funded by Naif Arab University for Security Sciences.

Institutional Review Board Statement: Ethical review and approval were waived or not required for this study because it used only publicly available secondary research benchmark datasets and did not involve direct interaction with human participants, intervention, recruitment of participants, collection of new human data, or access to private clinical or biomedical records by the authors. The experiments were conducted only on publicly released datasets, namely FakeAVCeleb and the DFDC, for secondary computational analysis of audio-visual deepfake detection. Ethics committee name: Not applicable. Approval code: not applicable. Approval date: not applicable.

Informed Consent Statement: Not applicable. This study used publicly available secondary datasets and did not involve recruitment of participants, intervention with human subjects, collection of new human data, or access to private clinical or biomedical records by the authors. The datasets were used in accordance with their respective terms and were cited in the manuscript.

Data Availability Statement: No new dataset was created in this study. The experiments used publicly available research benchmark datasets: FakeAVCeleb, available from the FakeAVCeleb project page at <https://sites.google.com/view/fakeavcelebdash-lab> (accessed on 16 July 2026), with an associated repository at <https://github.com/DASH-Lab/FakeAVCeleb> (accessed on 16 July 2026), and the Deepfake Detection Challenge (DFDC) dataset, available through the Kaggle challenge page at <https://www.kaggle.com/c/deepfake-detection-challenge> (accessed on 16 July 2026). The schematic elements in Figure 1 were redrawn by the authors or created from open-license graphical elements. Upon publication, the authors will release the implementation code, split files, raw prediction scores, threshold-sweep tables, FLOP and Brier analysis scripts, and component-ablation summaries in a public GitHub repository.

Acknowledgments: The authors thank Naif Arab University for Security Sciences for its institutional support and encouragement of this work.

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

References

- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. <https://doi.org/10.1145/3422622>.
- Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33; pp. 6840–6851.
- Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11. <https://doi.org/10.1109/ICCV.2019.00009>.
- Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C.Y.; Zhao, Y. Audio Deepfake Detection: A Survey. *arXiv* **2023**, arXiv:2308.14970. <https://doi.org/10.48550/arXiv.2308.14970>.
- Zhou, Y.; Lim, S.N. Joint Audio-Visual Deepfake Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14780–14789. <https://doi.org/10.1109/ICCV48922.2021.01453>.
- Khalid, H.; Kim, M.; Tariq, S.; Woo, S.S. 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, Virtual, 24 October 2021; pp. 7–15. <https://doi.org/10.1145/3476099.3484315>.
- Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The Deepfake Detection Challenge Dataset. *arXiv* **2020**, arXiv:2006.07397 <https://doi.org/10.48550/arXiv.2006.07397>.
- Chang, A.; Guzman, A.R.; Park, H.; Hopkins, K.; Yang, T.; Shrutika, S. 2D3MF: Deepfake Detection Using Multi-Modal Middle Fusion. Software Repository and Technical Manuscript. 2024. Available online: <https://github.com/aiden200/2D3MF> (accessed on 16 July 2026).
- Yang, J.; Li, C.; Dai, X.; Yuan, L.; Gao, J. Focal Modulation Networks. *arXiv* **2022**, arXiv:2203.11926. <https://doi.org/10.48550/arXiv.2203.11926>.
- Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5534–5542. <https://doi.org/10.1109/ICCV.2017.590>.
- Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459. <https://doi.org/10.1109/CVPR.2018.00675>.
- Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6804–6815. <https://doi.org/10.1109/ICCV48922.2021.00675>.
- Javed, M.; Zhang, Z.; Dahri, F.H.; Laghari, A.A.; Krajčák, M.; Almadhor, A. Real-Time Deepfake Detection via Gaze and Blink Patterns: A Transformer Framework. *Comput. Mater. Contin.* **2025**, *85*, 1457–1493. <https://doi.org/10.32604/cmc.2025.062954>.
- Javed, M.; Zhang, Z.; Dahri, F.H.; Laghari, A.A.; Krajčák, M.; Almadhor, A. Audio-Visual Synchronization and Lip Movement Analysis for Real-Time Deepfake Detection. *Int. J. Comput. Intell. Syst.* **2025**, *18*, 170. <https://doi.org/10.1007/s44196-025-00911-7>.
- Javed, M.; Zhang, Z.; Dahri, F.H.; Kumar, T. Enhancing Multimodal Deepfake Detection with Local-Global Feature Integration and Diffusion Models. *Signal Image Video Process.* **2025**, *19*, 400.
- Javed, M.; Zhang, Z.; Dahri, F.H.; Laghari, A.A. Real-Time Deepfake Video Detection Using Eye Movement Analysis with a Hybrid Deep Learning Approach. *Electronics* **2024**, *13*, 2947. <https://doi.org/10.3390/electronics13152947>.
- Javed, M.; Han, D.; Dahri, F.H.; Kumar, T.; Bhutto, J.A.; Khowaja, A. Real-Time Detection of Deepfakes Using the Efficient Swin Attention Network with Global and Local Facial Features. *Discov. Artif. Intell.* **2026**, *6*, 508. <https://doi.org/10.1007/s44163-026-01188-1>.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988. <https://doi.org/10.1109/CVPR52688.2022.01553>.
- Tong, Z.; Song, Y.; Wang, J.; Wang, L. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 10078–10093.
- Cai, Z.; Ghosh, S.; Stefanov, K.; Dhall, A.; Cai, J.; Rezatofighi, H.; Haffari, G.; Hayat, M. MARLIN: Masked Autoencoder for Facial Video Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1493–1504. <https://doi.org/10.1109/CVPR52729.2023.00150>.

21. Hamza, A.; Javed, A.R.; Iqbal, F.; Kryvinska, N.; Almadhor, A.S.; Jalil, Z.; Borghol, R. Deepfake Audio Detection via MFCC Features Using Machine Learning. *IEEE Access* **2022**, *10*, 134018–134028. <https://doi.org/10.1109/ACCESS.2022.3231480>.
22. Mcuba, M.; Singh, A.; Ikuesan, R.A.; Venter, H. The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation. *Procedia Comput. Sci.* **2023**, *219*, 211–219. <https://doi.org/10.1016/j.procs.2023.01.283>.
23. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 12449–12460.
24. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447. <https://doi.org/10.48550/arXiv.2106.07447>.
25. Baevski, A.; Hsu, W.N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Proceedings of Machine Learning Research; PMLR: 2022; Volume 162, pp. 1298–1312.
26. Gong, Y.; Lai, C.I.; Chung, Y.A.; Glass, J. SSAST: Self-Supervised Audio Spectrogram Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 10699–10709. <https://doi.org/10.1609/aaai.v36i10.21315>.
27. Huang, P.Y.; Xu, H.; Li, J.; Baevski, A.; Auli, M.; Galuba, W.; Metze, F.; Feichtenhofer, C. Masked Autoencoders that Listen. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 28708–28720.
28. Chen, W.; Liang, Y.; Ma, Z.; Zheng, Z.; Chen, X. EAT: Self-supervised Pre-training with Efficient Audio Transformer. *arXiv* **2024**, arXiv:2401.03497. <https://doi.org/10.48550/arXiv.2401.03497>.
29. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Inf. Fusion* **2020**, *64*, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>.
30. Müller, N.M.; Czempin, P.; Dieckmann, F.; Froghyar, A.; Böttinger, K. Does Audio Deepfake Detection Generalize? *arXiv* **2022**, arXiv:2203.16263. <https://doi.org/10.48550/arXiv.2203.16263>.
31. Chugh, K.; Gupta, P.; Dhall, A.; Subramanian, R. Not Made for Each Other: Audio-Visual Dissonance-Based Deepfake Detection and Localization. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 439–447. <https://doi.org/10.1145/3394171.3413700>.
32. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2823–2832. <https://doi.org/10.1145/3394171.3413570>.
33. Hosler, B.; Salvi, D.; Murray, A.; Antonacci, F.; Bestagini, P.; Tubaro, S.; Stamm, M.C. Do Deepfakes Feel Emotions? A Semantic Approach to Detecting Deepfakes via Emotional Inconsistencies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2021; pp. 1013–1022. <https://doi.org/10.1109/CVPRW53098.2021.00112>.
34. Conti, E.; Salvi, D.; Borrelli, C.; Hosler, B.; Bestagini, P.; Antonacci, F.; Sarti, A.; Stamm, M.C.; Tubaro, S. Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 8962–8966. <https://doi.org/10.1109/ICASSP43922.2022.9747186>.
35. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>.
36. Oorloff, T.; Koppiseti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; Bharaj, G. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. *arXiv* **2024**, arXiv:2406.02951. <https://doi.org/10.48550/arXiv.2406.02951>.
37. Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; Ren, K. AVoid-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 2015–2029. <https://doi.org/10.1109/TIFS.2023.3262148>.
38. Wang, R.; Ye, D.; Tang, L.; Zhang, Y.; Deng, J. AVT²-DWF: Improving Deepfake Detection with Audio-Visual Fusion and Dynamic Weighting Strategies. *arXiv* **2024**, arXiv:2403.14974. <https://doi.org/10.48550/arXiv.2403.14974>.
39. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3204–3213. <https://doi.org/10.1109/CVPR42600.2020.00327>.
40. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed.; Springer: New York, NY, USA, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
41. Liu, P.; Tao, Q.; Zhou, J.T. Evolving from Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges. *arXiv* **2024**, arXiv:2406.06965. <https://doi.org/10.48550/arXiv.2406.06965>.

42. Astrid, M.; Ghorbel, E.; Aouada, D. Statistics-aware Audio-visual Deepfake Detector. *arXiv* **2024**, arXiv:2407.11650. <https://doi.org/10.48550/arXiv.2407.11650>.
43. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*; Proceedings of Machine Learning Research, PMLR: New York, NY, USA, 2017; Volume 70, pp. 3319–3328.

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