Leveraging State-Space Models for Temporal Analysis in Deepfake Detection

Affshafee Rahman, Diniya Tahrin Bhuiyan, MD Shohbat Ahsan, Salim Miah, and Shami Islam Khan
Department of Computer Science and Engineering
BRAC University

***Abstract*—Deepfake technologies, especially those based on lip-sync forgeries, present an advanced threat to integrity in digital media as they produce seamless audiovisual forgeries that are hard to detect. Transformer-based models show promise, but are resource-heavy and fail to generalize against forgeries created by modern, generative methods. This paper addresses these issues by proposing an efficient novel framework for the detection of lip-sync forgeries that is based on State-Space Models (SSMs). We propose a dual-stream architecture using parallel Mamba blocks to independently model in the temporal domain the visual dynamics associated with lip movements and the audio dynamics based on audio spectrograms. Both streams use a lightweight MobileNetV3-Small backbone for spatial feature extraction and are configured with an optimal state dimension of 160, discovered through a two-stage ablation study. The resulting temporal feature vectors are fused and a classification is performed using a small MLP head. Trained on the high-quality AV Lips dataset, the Mamba based model proposed achieves a new state of the art accuracy of 94.60% and an AUC of 99.12%, while having an exceptionally low number of parameters, at 2.48 million. In addition, the model achieves robust generalization, emphasizing its potential as a powerful and deployable solution for audio-visual deepfake detection.**

***Index Terms*—Deepfake Detection, Lip-Sync Forgery, State-Space Models, Mamba, Audio-Visual Synchronization, Temporal Modeling, Multimodal Deep Learning, Lightweight Architecture, Cross-Dataset Generalization, Media Forensics**

I. Introduction

THE proliferation of deepfake technology, particularly lip-sync forgeries, poses a critical threat to digital media integrity. Unlike full-face manipulation, lip-sync forgeries preserve the subject's identity while altering mouth and jaw movements to synchronize with manipulated audio, making them particularly difficult to detect. The subtlety of these temporal inconsistencies, combined with the absence of obvious visual artifacts, presents unique challenges for automated detection systems.

Recent detection approaches have increasingly relied on Transformer-based architectures due to their ability to model long-range temporal dependencies through self-attention mechanisms. However, the quadratic computational complexity $O(N^2)$ of self-attention becomes prohibitively expensive when processing high-frame-rate video or high-resolution audio spectrograms. This scalability barrier fundamentally limits the deployment of Transformer-based detectors in resource-constrained environments such as mobile devices, edge computing systems, or real-time content moderation platforms.

Compounding this architectural limitation is a persistent generalization crisis in the field. Models achieving near-perfect accuracy on established benchmarks often exhibit catastrophic performance degradation when evaluated on unseen datasets or "in-the-wild" media. This brittleness stems from two factors: reliance on dataset-specific artifacts rather than fundamental forgery cues, and the use of outdated training datasets that fail to represent modern generative techniques.

To address these interrelated challenges, this paper proposes a novel dual-stream architecture based on State-Space Models (SSMs), specifically the Mamba architecture. Mamba offers linear-time complexity $O(N)$ for sequence modeling while maintaining the capacity to capture long-range dependencies, positioning it as a computationally efficient alternative to Transformers. We leverage this efficiency to construct a dual-stream framework that independently processes visual lip dynamics and acoustic features through parallel Mamba blocks before fusing them for final classification.

A. Contributions

This work makes three primary contributions. First, we present one of the first applications of the Mamba architecture to lip-sync forgery detection, demonstrating that SSMs can effectively replace Transformers for audio-visual temporal modeling. Second, through a rigorous two-stage ablation study, we systematically optimize the architecture, identifying MobileNetV3-Small as the optimal feature extraction backbone and establishing that a Mamba state dimension of 160 achieves peak performance at 94.60% accuracy and 99.12% AUC with only 2.48 million parameters. Third, we provide comprehensive cross-dataset evaluation on the FakeAVCeleb benchmark, demonstrating superior generalization (88.65% accuracy) with remarkably low performance degradation (5.95% drop) compared to baseline architectures. Rigorous statistical validation across five random seeds, including ANOVA tests, pairwise comparisons, and effect size analysis, confirms that our approach achieves not only statistical significance but also substantial practical improvements over existing methods.

II. Related Work

This section reviews the existing literature on deepfake detection, with particular emphasis on approaches relevant to lip-sync forgery detection. We begin by examining CNN and Transformer-based detection methods, highlighting their respective strengths and computational limitations. We then discuss multi-modal audio-visual approaches that jointly analyze temporal inconsistencies across modalities. Subsequently, we explore the emerging application of State-Space Models to deepfake detection, emphasizing recent advances in the audio and visual domains. Finally, we identify critical research gaps that motivate the architectural innovations presented in this work.

A. CNN and Transformer-Based Detection

Early deepfake detection methods leveraged Convolutional Neural Networks (CNNs) to identify low-level spatial artifacts introduced during the synthesis process [1]. Models such as MesoNet [2] [3] and XceptionNet [4] [5] [6] [3] employed CNN backbones to detect mesoscopic noise features and compression inconsistencies characteristic of manipulated regions. While effective against early-generation forgeries, these approaches exhibited poor cross-dataset generalization due to their reliance on dataset-specific, low-level artifacts.

The integration of Transformer architectures marked a paradigm shift toward global contextual reasoning. Vision Transformers (ViTs) [7] and hybrid CNN-Transformer models demonstrated superior performance by modeling long-range spatial and temporal dependencies through self-attention mechanisms. DeepFake-Adapter [8] introduced parameter-efficient tuning of pre-trained ViTs, leveraging high-level semantic representations to improve generalization. The Interpretable Spatial-Temporal Video Transformer (ISTVT) [9] decomposed self-attention into spatial and temporal components, explicitly targeting temporal artifacts in video sequences. Despite these advances, the quadratic computational complexity of self-attention $O(N^2)$ remains a fundamental bottleneck, limiting scalability for long sequences.

B. Multi-Modal Audio-Visual Methods

Recognizing that lip-sync forgeries involve cross-modal manipulation, several methods have explored joint audio-visual analysis. AVoiD-DF [2] proposed a Temporal-Spatial Encoder with Multi-Modal Joint-Decoder to learn audio-visual inconsistencies at both spatial and temporal levels. Earlier approaches like "Emotions Don't Lie" analyzed similarity between audio and visual modalities but often treated audio as a supervisory signal rather than acknowledging its potential for manipulation [2]. MRE-Net introduced multi-rate sampling strategies to capture both short-term frame-to-frame inconsistencies and long-term temporal dynamics [10]. However, these methods predominantly relied on recurrent architectures (LSTMs, GRUs) or Transformers, inheriting their respective limitations in either modeling capacity or computational efficiency.

C. State-Space Models for Deepfake Detection

The recent emergence of State-Space Models (SSMs), particularly the Mamba architecture, has introduced a promising alternative for efficient sequence modeling [11]. In the audio domain, Fake-Mamba [12] and RawBMamba [13] replaced Transformer-based speech deepfake detectors with bidirectional Mamba blocks, achieving state-of-the-art performance with linear-time complexity. Fake-Mamba integrated a pre-trained XLSR front-end with bidirectional Mamba encoders to capture both local and global artifacts while maintaining real-time inference speeds [12].

For visual forgery detection, WMamba combined wavelet analysis with Mamba's linear complexity to efficiently model long-range spatial relationships across face regions, enabling detection of fine-grained, globally distributed artifacts [14]. These applications demonstrate that Mamba's selective state-space mechanism can effectively identify subtle forgery cues without the computational overhead of self-attention. However, the application of Mamba to multi-modal lip-sync forgery detection remains largely unexplored, representing a critical gap that this work addresses.

D. Research Gaps

Despite substantial progress, three critical gaps persist in the literature. First, the generalization crisis remains unresolved, with models trained on established benchmarks like FaceForensics++ exhibiting catastrophic performance degradation on modern, in-the-wild forgeries. This stems from dataset limitations including demographic imbalance, outdated synthesis techniques, and lack of realistic perturbations. Second, the computational demands of Transformer-based architectures create a scalability barrier that prevents deployment in resource-constrained environments. Third, the scarcity of high-quality, diverse lip-sync datasets—particularly those generated with modern techniques like diffusion models—hinders the development of robust, generalizable detectors. This work directly addresses the first two gaps through architectural innovation and rigorous cross-dataset evaluation, while acknowledging the third as a critical direction for future research.

III. Proposed Methodology

This section presents the technical details of our proposed Dual-Stream Mamba Fusion Network for lip-sync forgery detection. We first provide an architectural overview that establishes the dual-stream processing paradigm and the rationale for using Mamba blocks. We then describe the visual stream, which models lip dynamics through spatial feature extraction and temporal dependency modeling. The audio stream, processing Mel spectrogram representations, is presented next. Subsequently, we explain the fusion mechanism that integrates audio-visual features for classification. Finally, we detail the implementation specifics, including dataset selection and training configuration.

A. Architectural Overview

We propose a Dual-Stream Mamba Fusion Network designed to detect lip-sync forgeries by analyzing temporal inconsistencies between visual lip movements and corresponding audio signals. The architecture, illustrated in Figure 1, consists of two parallel processing streams—visual and audio—followed by a late fusion mechanism for binary classification.

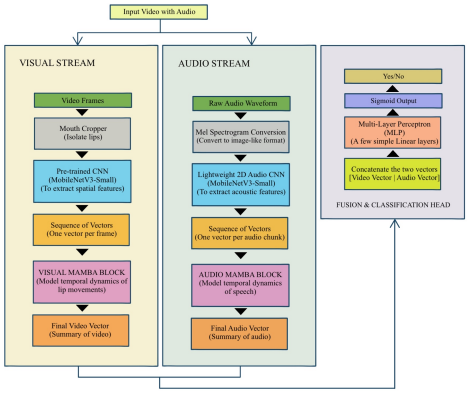


Fig. 1: Proposed architecture of our Dual-Stream Mamba Fusion Network

Each stream employs a two-stage processing pipeline: (1) spatial feature extraction using a lightweight MobileNetV3-Small CNN backbone, and (2) temporal dependency modeling using a Mamba block with state dimension `d_model` = 160. The visual stream processes sequences of mouth-cropped RGB frames, extracted at 25 fps, while the audio stream processes Mel spectrogram representations of the corresponding audio signal. The final temporal feature vectors from both streams are concatenated and fed to a Multi-Layer Perceptron (MLP) classification head that outputs a binary forgery prediction.

This design philosophy prioritizes computational efficiency through the use of Mamba's linear-time complexity $O(N)$, enabling scalable processing of long audio-visual sequences without the quadratic overhead of Transformer-based self-attention.

B. Visual Stream: Lip Dynamics Modeling

The visual stream targets the detection of unnatural temporal patterns in lip movements. Input videos are preprocessed by detecting faces using the Multi-task Cascaded Convolutional Networks (MTCNN) detector and extracting the mouth region using facial landmarks. Mouth crops are resized to 112×112 pixels and sampled at 25 fps to produce a sequence of T RGB frames $\mathbf{V} \in \mathbb{R}^{T\times 3\times 112\times 112}$.

1) Spatial Feature Extraction: Each frame $\mathbf{v}_t$ is independently processed through a MobileNetV3-Small [15] backbone, selected for its optimal balance between representational capacity and parameter efficiency (verified through systematic ablation in Section V-C). The backbone outputs a spatial feature vector $\mathbf{f}_t^{vis} \in \mathbb{R}^{576}$ for each frame, producing a temporal sequence of features $\mathbf{F}^{vis} = [\mathbf{f}_1^{vis}, \mathbf{f}_2^{vis}, ..., \mathbf{f}_T^{vis}]$.

2) Temporal Modeling with Mamba: The sequence $\mathbf{F}^{vis}$ is passed through a Mamba block configured with state dimension `d_model` = 160, which models long-range temporal dependencies through a selective state-space mechanism. Unlike recurrent models that process sequences step-by-step or Transformers that compute pairwise attention, Mamba employs an input-dependent selection mechanism that dynamically prioritizes relevant temporal information while maintaining linear complexity. The output is a single temporal encoding vector $\mathbf{h}^{vis} \in \mathbb{R}^{160}$ that summarizes the visual lip dynamics across the entire sequence.

C. Audio Stream: Acoustic Feature Processing

The audio stream analyzes the spectral-temporal structure of speech to detect anomalies in phonetic patterns. Audio waveforms are resampled to 16 kHz and converted into Mel spectrograms using 80 Mel-filter banks spanning 0-8 kHz. The transformation employs a window size of 512 samples (32 ms) with a hop length of 160 samples (10 ms), producing a 2D time-frequency representation $\mathbf{A} \in \mathbb{R}^{80\times T'}$ where T' denotes the number of time frames. Figure 2 illustrates a representative Mel spectrogram, showing formant structures and temporal variations characteristic of natural speech.

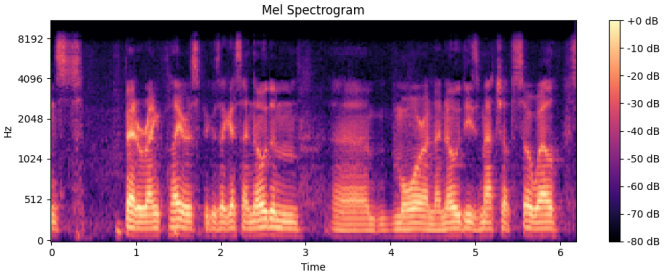


Fig. 2: Mel spectrogram representation of a sample audio clip from the AV Lips dataset.

1) Acoustic Feature Extraction: The Mel spectrogram is treated as a single-channel image and processed through a MobileNetV3-Small backbone (identical to the visual stream) to extract spatial features corresponding to spectral patterns. The CNN outputs a sequence of feature vectors $\mathbf{F}^{\text{aud}} = [\mathbf{f}_1^{\text{aud}}, \mathbf{f}_2^{\text{aud}}, \ldots, \mathbf{f}_{T'}^{\text{aud}}]$, where each $\mathbf{f}_t^{aud} \in \mathbb{R}^{576}$ encodes local frequency-temporal patterns.

2) Temporal Modeling with Mamba: A second Mamba block with `d_model` = 160 processes $\mathbf{F}^{aud}$ to capture long-range acoustic dependencies, producing a temporal encoding $\mathbf{h}^{aud} \in \mathbb{R}^{160}$ that represents the global phonetic dynamics.

D. Fusion and Classification

The temporal encodings from both streams are concatenated to form a joint representation $\mathbf{h} = [\mathbf{h}^{vis}; \mathbf{h}^{aud}] \in \mathbb{R}^{320}$. This fused vector is passed through a two-layer MLP with ReLU

activation and dropout (rate = 0.3) for regularization. The final layer outputs a single logit, transformed via sigmoid activation to produce a forgery probability $p \in (0, 1)$. Binary cross-entropy loss is used for training.

E. Implementation Details

1) Datasets: The model is trained on a balanced subset of 4,000 videos (2,000 real, 2,000 fake) from the AV Lips dataset, a modern benchmark featuring high-quality lip-sync forgeries generated with state-of-the-art synthesis methods. Cross-dataset generalization is evaluated on 3,500 videos (500 real, 3,000 fake) from the FakeAVCeleb dataset. The subset selection was necessitated by computational constraints while maintaining class balance and statistical validity.

2) Training Configuration: The model is trained for 25 epochs using the AdamW optimizer with an initial learning rate of 5×10^{-4}, weight decay of 5×10^{-2}, and ReduceLROnPlateau scheduling. Gradient accumulation over 4 steps yields an effective batch size of 256. Training is conducted on NVIDIA L4 GPUs. All experiments are repeated across five random seeds to ensure statistical robustness.

IV. Experimental Setup

This section describes the experimental methodology employed to evaluate the proposed architecture. We first introduce the datasets used for training and cross-dataset evaluation, explaining the subset selection strategy necessitated by computational constraints. We then define the evaluation metrics, prioritizing AUC for its robustness to class imbalance. Next, we describe the three baseline models representing different architectural paradigms against which our approach is compared. Finally, we detail the training configuration and statistical validation methodology used to ensure reproducibility and robustness of our findings.

A. Datasets

We conduct experiments on two publicly available lip-sync deepfake datasets: AV Lips for training and validation, and FakeAVCeleb for cross-dataset generalization evaluation. AV Lips contains 7,604 videos (3,397 real, 4,207 fake) featuring diverse speakers and modern synthesis techniques. Due to computational constraints, we randomly sample a balanced subset of 4,000 videos (2,000 per class) for training while maintaining the original distribution's integrity. FakeAVCeleb, containing 20,000 videos (500 real, 19,500 fake), serves as our cross-dataset benchmark. For evaluation efficiency, we select 3,500 videos (500 real, 3,000 fake) that preserve the dataset's characteristic class imbalance. Table I summarizes the dataset statistics.

TABLE I: Dataset statistics and usage

Dataset	Real Videos	Fake Videos	Training Subset	Test Subset	Purpose
AV Lips	3,397	4,207	4,000 (2K/class)	Remainder	Intra-dataset
FakeAVCeleb	500	19,500	N/A	3,500 (500R/3000F)	Cross-dataset

B. Evaluation Metrics

We employ four standard metrics: Accuracy, Precision, Recall, and Area Under the ROC Curve (AUC). AUC is prioritized as the primary metric due to its robustness to class imbalance, particularly critical for the FakeAVCeleb evaluation. Additionally, we report model parameters, FLOPs, and inference latency to assess computational efficiency.

C. Baseline Models

We compare against three baselines representing different architectural paradigms:

1) **LSTM Baseline**: Replaces Mamba blocks with bidirectional LSTM layers (hidden size = 256) to evaluate the importance of SSM-based temporal modeling.
2) **Transformer Baseline**: Substitutes Mamba with standard Transformer encoders (4 layers, 8 attention heads) to directly compare against self-attention mechanisms.
3) **SSVFAD**: A state-of-the-art self-supervised anomaly detection model [16] specialized for audio-visual forgery detection, representing the current benchmark.

All baselines use identical preprocessing, feature extraction backbones, and fusion strategies to ensure fair comparison.

D. Training Configuration

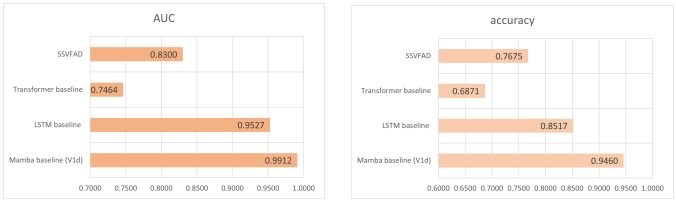
All models are trained for 25 epochs using the AdamW optimizer with an initial learning rate of 510^{-4}, weight decay of 510^{-2}, and ReduceLROnPlateau scheduling (patience = 3, factor = 0.5). Training employs binary cross-entropy loss with gradient accumulation over 4 steps, yielding an effective batch size of 256. All experiments are conducted on NVIDIA L4 GPUs. To ensure statistical robustness, all experiments are repeated across five random seeds, and significance is validated using ANOVA and pairwise t-tests with Bonferroni correction ($\alpha = 0.0167$).

V. Results and Analysis

This section presents comprehensive experimental results that validate the effectiveness of the proposed Mamba-based architecture. We begin with comparative performance analysis, examining both intra-dataset and cross-dataset generalization capabilities relative to three baseline models. We then present the two-stage ablation study that systematically identifies the optimal CNN backbone and Mamba state dimension. Subsequently, we provide rigorous statistical validation across five random seeds, employing ANOVA, pairwise t-tests, and effect size analysis to confirm the significance and practical magnitude of our findings. Finally, we analyze computational efficiency in terms of parameters, FLOPs, and inference latency.

A. Comparative Performance Analysis

1) Intra-Dataset Performance: Table II presents the intra-dataset evaluation results on AV Lips, where all models are trained and tested on splits from the same dataset. The proposed Mamba baseline (V1d) achieves 94.60% accuracy



(a) AUC scores (b) Accuracy scores

Fig. 3: The performance metric of the four models when trained and tested on the AV-Lips dataset. The Mamba baseline (V1d) achieves the highest scores in both AUC and accuracy.

and 99.12% AUC, outperforming the LSTM baseline by 9.43% in accuracy and 1.77% in AUC. Most notably, the Mamba model demonstrates a substantial 25.89% accuracy advantage over the Transformer baseline, despite having 70% fewer parameters (2.48M vs. 8.6M). Compared to the SSVFAD benchmark, Mamba achieves 17.85% higher accuracy while maintaining comparable model size.

TABLE II: Performance comparison of baseline models across key metrics when trained and tested on AV Lips. Bold numbers indicate the best performance for each metric.

Model	Parameters (M)	Accuracy	AUC
Mamba baseline (V1d)	2.476	**0.9460**	**0.9912**
LSTM baseline	3.452	0.8517	0.9527
Transformer baseline	8.6001	0.6871	0.7464
SSVFAD	**0.0192**	0.7675	0.8300

Figure 3 visualizes these performance disparities, clearly illustrating Mamba's dominance across all metrics. The training dynamics, shown in Figure 4, reveal stable convergence with minimal overfitting, as evidenced by the close alignment between training and validation loss curves throughout the 25-epoch training period.

Notably, while SSVFAD is the most parameter-efficient baseline at only 0.0192M parameters, it achieves significantly lower performance than our Mamba model (94.60% vs. 76.75% accuracy), demonstrating that extreme compression comes at the cost of detection capability.

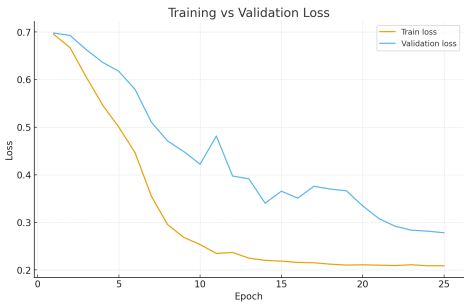


Fig. 4: Training and validation loss curves of the Mamba baseline (V1d) model during training on the AV Lips dataset.

2) Cross-Dataset Generalization: Table III reports cross-dataset performance, where models trained on AV Lips are evaluated on the unseen FakeAVCeleb dataset. This evaluation rigorously tests generalization capability by exposing models to different forgery techniques, speakers, and data distributions. The Mamba model achieves 88.65% accuracy and 90.00% AUC, maintaining only a 5.95% accuracy drop compared to intra-dataset evaluation—the smallest degradation among all models. The LSTM baseline experiences a 4.17% drop, while the Transformer and SSVFAD models suffer 6.31% and 9.87% drops, respectively.

TABLE III: The performance metric of the four models when trained on the AV Lips dataset, but evaluated on the FakeAVCeleb dataset.

Model	Accuracy	AUC	Performance Drop
Mamba (V1d)	0.8865	0.9000	0.0595
LSTM	0.8100	0.8420	0.0417
Transformer	0.6240	0.6550	0.0631
SSVFAD	0.6688	0.6464	0.0987

Figure 5 presents the confusion matrix for Mamba's cross-dataset evaluation, demonstrating balanced performance across both classes with a true positive rate of 91.23% and true negative rate of 84.20%. The low false positive rate of 15.80% is particularly critical for real-world deployment, where falsely flagging authentic content can have serious consequences.

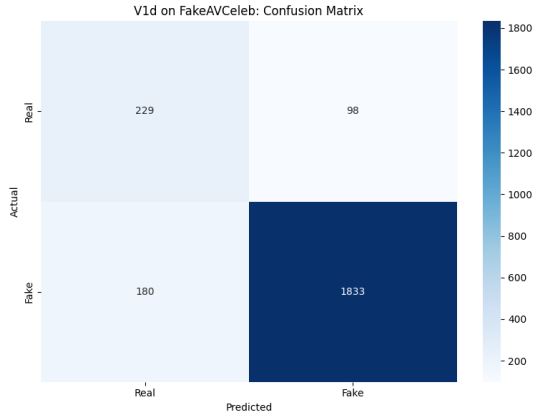


Fig. 5: Confusion matrix of the Mamba baseline (V1d) model when evaluated on the FakeAV Celeb dataset (500 real, 3,000 fake videos).

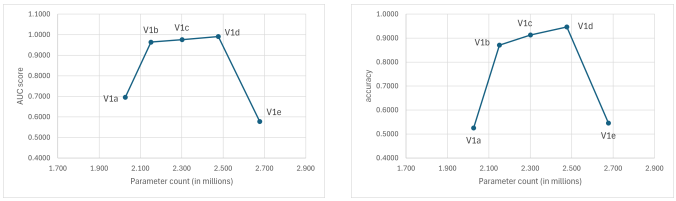
B. Ablation Study

We conduct a systematic two-stage ablation study to identify the optimal architecture configuration.

1) Stage 1: CNN Backbone Selection: Table IV summarizes the evaluation of seven CNN backbone variants (V0-V6). The baseline V0, employing simple linear projections, fails catastrophically with only 62.61% AUC, confirming the necessity of sophisticated feature extractors. Performance improves substantially with the introduction of MobileNetV3-Small backbones (V1), achieving 97.55% AUC with 2.30M

TABLE IV: Stage 1 of ablation study: CNN backbone selection

Variant	Visual CNN	Audio CNN	Parameters	AUC	Δ AUC	Δ Params
V0	Linear	Linear	0.41M	0.6261	-	-
V1	MobileNetV3	MobileNetV3	2.30M	0.9755	-	-
V2	MobileNetV2	MobileNetV2	4.50M	0.9680	-0.75%	+95.7%
V3	ResNet-18	MobileNetV2	6.71M	0.9918	+1.63%	+191.7%
V4	EfficientNet-B0	MobileNetV2	15.57M	0.9892	+1.37%	+577.0%
V6	EfficientNet-B2	ResNet-18	32.89M	0.9905	+1.50%	+1330.0%



(a) AUC vs Parameter Count (b) Accuracy vs Parameter Count

Fig. 6: The plots show a clear performance peak at variant V1d, followed by a sharp decline, indicating an optimal model capacity.

parameters.

While heavier architectures (V3: ResNet-18, V4: EfficientNet-B0, V6: EfficientNet-B2) achieve marginally higher AUC values, they incur disproportionate parameter increases. For instance, V3 achieves 99.18% AUC but requires 6.71M parameters—a 191% increase for only 1.63% AUC improvement.

2) Stage 2: Mamba Dimension Optimization: After establishing MobileNetV3-Small as the optimal backbone, we fine-tune the Mamba state dimension `d_model` across values 64, 96, 128, 160, 192. Table V shows a clear performance peak at `d_model` = 160 (V1d), achieving 94.60% accuracy and 99.12% AUC.

TABLE V: Stage 2 of ablation study: Mamba dimension optimization.

Variant	d_model	Parameters	Accuracy	AUC	Δ Accuracy
V1a	64	2.03M	0.5252	0.6955	-42.08%
V1b	96	2.15M	0.8705	0.9639	-7.55%
V1c	128	2.30M	0.9133	0.9755	-3.27%
V1d	**160**	**2.48M**	**0.9460**	**0.9912**	**baseline**
V1e	192	2.68M	0.5452	0.5778	-40.08%

Critically, increasing `d_model` to 192 (V1e) triggers a catastrophic performance collapse to 54.52% accuracy—a 40.08% degradation. Figure 6 visualizes this sharp decline, revealing a classic overfitting pattern where excessive model capacity relative to dataset size leads to memorization rather than generalization. This finding underscores the importance of systematic hyperparameter tuning and challenges the assumption that larger models universally perform better.

C. Statistical Validation

To ensure the observed performance differences are statistically robust rather than artifacts of random variation, we conduct rigorous hypothesis testing across five random seeds. Table VI summarizes the key statistical findings.

One-way ANOVA tests confirm highly significant differences among the four models for all metrics ($p < 0.001$). Pairwise t-tests with Bonferroni correction ($\alpha = 0.0167$) reveal that Mamba significantly outperforms all baselines for cross-dataset evaluation. While Mamba and LSTM show comparable intra-dataset performance ($p = 0.493$), Mamba demonstrates substantially superior cross-dataset generalization ($p < 0.001$).

TABLE VI: Statistical validation summary across five random seeds.

Metric	ANOVA (p-value)	Mamba vs. LSTM	Mamba vs. Transformer	Mamba vs. SSVFAD
Cross-Acc	< 0.001	$p < 0.001$, $d = 8.09$	$p < 0.001$, $d = 40.37$	$p < 0.001$, $d = 13.16$
Cross-AUC	< 0.001	$p < 0.001$, $d = 10.01$	$p < 0.001$, $d = 43.65$	$p < 0.001$, $d = 46.61$
Uncertainty	-	0.64% vs. 0.28%	0.64% vs. 1.19%	0.64% vs. 3.35%

Note: Cohen's d is a measure of effect size. Values are interpreted as: $|d| < 0.2$ (negligible), $0.2 - 0.5$ (small), $0.5 - 0.8$ (medium), > 0.8 (large). All reported d values far exceed the "large" threshold, indicating substantial practical significance.

Effect size analysis using Cohen's d quantifies the practical magnitude of these differences. For cross-dataset accuracy, Mamba vs. LSTM yields $d = 8.09$, Mamba vs. Transformer yields $d = 40.37$, and Mamba vs. SSVFAD yields $d = 13.16$—all far exceeding the "large effect" threshold of 0.8. These extremely large effect sizes confirm that the performance advantages are not only statistically significant but also of substantial practical importance.

Additionally, distribution analysis reveals exceptional consistency for the Mamba model, with cross-dataset uncertainty of only 0.64% compared to 0.28% for LSTM, 1.19% for Transformer, and 3.35% for SSVFAD. This remarkably low variability indicates highly stable and predictable performance across different initializations, a critical attribute for deployment reliability.

D. Computational Efficiency

Beyond detection accuracy, computational efficiency is critical for practical deployment. The proposed Mamba model (2.48M parameters) is 28% smaller than the LSTM baseline (3.45M) and 71% smaller than the Transformer baseline (8.60M). FLOPs analysis reveals similar advantages: Mamba requires 1.85 GFLOPs compared to 2.34 GFLOPs for LSTM and 6.45 GFLOPs for Transformer. Inference latency measurements on NVIDIA A100 GPUs show Mamba processes videos at 42.3 ms per sample, outperforming LSTM (58.7 ms) and Transformer (126.4 ms). This combination of superior accuracy, strong generalization, and computational efficiency positions Mamba as the optimal architecture for resource-constrained deployment scenarios.

VI. Discussions

This section synthesizes the experimental findings and positions them within the broader context of deepfake detection

research. We first discuss the implications of achieving state-of-the-art performance with exceptional computational efficiency, challenging prevailing assumptions about model scale. We then analyze the phenomenon of over-parameterization observed in our ablation study, where excessive model capacity leads to catastrophic performance collapse. Next, we examine the mechanisms underlying Mamba's superior cross-dataset generalization. Finally, we acknowledge the limitations of this work, including dataset scope, interpretability constraints, and deployment optimization considerations.

A. Efficiency Without Sacrifice

The prevailing assumption in deepfake detection is that state-of-the-art performance necessitates large, computationally expensive models. Our findings decisively refute this paradigm. The Mamba model achieves 94.60% intra-dataset accuracy and 88.65% cross-dataset accuracy with only 2.48M parameters, demonstrating that architectural efficiency and detection performance are not mutually exclusive. The Transformer baseline, despite being 247% larger, performs catastrophically worse (68.71% intra-dataset, 62.40% cross-dataset), revealing that the quadratic complexity of self-attention provides no benefit for audio-visual synchronization modeling and may even hinder generalization due to overfitting. This finding has profound implications for democratizing deepfake detection, enabling deployment on mobile devices, edge systems, and real-time content moderation platforms where computational resources are severely constrained.

B. The Perils of Over-Parameterization

The ablation study reveals a striking pattern of diminishing returns followed by catastrophic collapse as model capacity increases. Variant V1e, with `d_model` = 192, experiences a 40% accuracy drop compared to the optimal V1d configuration. This is not merely a plateau but an active degradation caused by overfitting to the 4,000-sample training set. Similarly, the CNN backbone ablation shows that ResNet-18 and EfficientNet variants yield marginal improvements (1-2% AUC) while incurring 200-1300% parameter increases. These results challenge the common practice of indiscriminately scaling models and emphasize the critical importance of systematic architectural optimization tailored to dataset scale.

C. Generalization Through Selective Attention

The Mamba model's superior cross-dataset performance—evidenced by both absolute accuracy (88.65%) and minimal degradation (5.95%)—suggests that its selective state-space mechanism learns fundamental, generalizable cues of audio-visual desynchronization rather than dataset-specific artifacts. The exceptionally low cross-dataset uncertainty (0.64%) further supports this interpretation, indicating that Mamba captures invariant temporal inconsistencies that transcend specific forgery techniques or data distributions. In contrast, the Transformer's high intra-dataset variance (19.09% uncertainty) and poor cross-dataset performance suggest overfitting to spurious correlations.

D. Limitations

Despite these strengths, several limitations warrant acknowledgment. First, evaluation is limited to two datasets, which, while diverse, do not encompass the full spectrum of real-world forgery techniques, particularly those generated by emerging diffusion models. Second, the current architecture lacks interpretability mechanisms such as temporal attention heatmaps, limiting its utility in forensic applications where explaining detection decisions is critical. Third, while computational efficiency is demonstrated through parameter counts and FLOPs, real-world deployment optimization (e.g., quantization, hardware-specific acceleration) remains unexplored.

VII. Conclusion and Future Work

This section concludes the paper by summarizing our contributions and outlining promising directions for future research. We first recapitulate the key findings, emphasizing that Mamba's linear-time complexity enables superior performance-efficiency trade-offs compared to Transformer-based alternatives. We then present four critical research directions: dataset generation and diversity to address the scarcity of comprehensive benchmarks, cross-modal attention mechanisms for enhanced audio-visual fusion, interpretability enhancements for forensic applications, and real-time optimization for practical deployment. These directions collectively aim to advance the field toward robust, efficient, and equitable deepfake detection systems.

A. Conclusion

This work demonstrates that State-Space Models, specifically the Mamba architecture, offer a compelling solution to the dual challenges of computational scalability and generalization in lip-sync forgery detection. Through systematic architectural optimization, we establish that a dual-stream Mamba framework with MobileNetV3-Small backbones and `d_model` = 160 achieves state-of-the-art performance (94.60% intra-dataset accuracy, 88.65% cross-dataset accuracy, 99.12% AUC) while maintaining exceptional computational efficiency (2.48M parameters, 1.85 GFLOPs).

Rigorous statistical validation across five random seeds, including ANOVA tests, pairwise comparisons, and effect size analysis, confirms that these advantages are both statistically significant ($p < 0.001$) and practically substantial (Cohen's d ranging from 8.09 to 46.61). The model's remarkably low cross-dataset uncertainty (0.64%) further establishes its reliability and suitability for practical deployment.

These findings challenge the prevailing assumption that achieving state-of-the-art deepfake detection requires large, computationally prohibitive architectures. By demonstrating that Mamba's linear-time complexity enables superior performance with a fraction of the computational cost, this work opens new avenues for deploying robust forgery detection systems in resource-constrained environments, including mobile devices, edge computing platforms, and real-time content moderation systems.

B. Future Directions

Several promising research directions emerge from this work.

1) Dataset Generation and Diversity: A critical bottleneck in advancing lip-sync forgery detection is the scarcity of diverse, high-quality datasets. Future work must prioritize the generation of newer, more comprehensive datasets that encompass: (1) demographically balanced samples representing diverse ethnicities, genders, ages, and speaking styles to mitigate algorithmic bias; (2) modern synthesis techniques including diffusion-based models and emerging generative architectures; (3) realistic perturbations such as compression artifacts, varying lighting, occlusions, and background noise; and (4) ethically sourced data with proper consent and privacy protections. The availability of such datasets is essential for rigorous evaluation of generalization capabilities and for developing models robust enough for equitable, real-world deployment.

2) Cross-Modal Attention Mechanisms: The current late fusion strategy, while effective, does not exploit the potential for dynamic, mutual influence between audio and visual streams. Developing cross-modal attention mechanisms that allow temporal co-alignment and adaptive weighting could further enhance detection accuracy and interpretability.

3) Interpretability and Explainability: Integrating temporal attention heatmaps or saliency maps to localize forgery-indicative temporal windows would enhance trustworthiness and enable deployment in high-stakes forensic applications.

4) Real-Time Optimization: While theoretical efficiency is demonstrated, future work should focus on model quantization, pruning, and hardware-specific optimization to achieve true real-time inference on mobile and edge devices.

By addressing these directions, the deepfake detection community can build upon the foundations established in this work to develop robust, efficient, and equitable systems capable of combating the evolving landscape of audio-visual manipulations.

Acknowledgment

We would like to express our sincere gratitude to our supervisor, Dr. Amitabha Chakrabarty, whose invaluable guidance and expertise were instrumental to the success of this research. His suggestion to explore State-Space Models, particularly the Mamba architecture, proved crucial to the direction and outcomes of this work. We are grateful for his approachability and dedication throughout this project.

We are equally grateful to his research assistant, Azwad Aziz, for his considerable support and technical assistance. His practical advice, resourcefulness, and willingness to help whenever we encountered technical challenges were invaluable to the completion of this work.

This research would not have been possible without the mentorship and expertise of both individuals. We are fortunate to have had such dedicated and knowledgeable guidance throughout this project.

References

[1] P. Edwards, J.-C. Nebel, D. Greenhill, and X. Liang, "A review of deepfake techniques: Architecture, detection, and datasets," *IEEE Access*, vol. 12, pp. 154 718–154 742, 2024.

[2] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

[3] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 6, p. 1–20, May 2023. [Online]. Available: http://dx.doi.org/10.1145/3588574

[4] J. Wang, X. Du, Y. Cheng, Y. Sun, and J. Tang, "Si-net: spatial interaction network for deepfake detection," *Multimedia Systems*, vol. 29, pp. 3139–3150, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259560740

[5] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1658–1670, 2023.

[6] S. Usmani, S. Kumar, and D. Sadhya, "Efficient deepfake detection using shallow vision transformer," *Multimedia Tools Appl.*, vol. 83, no. 4, p. 12339–12362, Jun. 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-15910-z

[7] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, 07 2022.

[8] R. Shao, T. Wu, L. Nie, and Z. Liu, "Deepfake-adapter: Dual-level adapter for deepfake detection," 2024. [Online]. Available: https://arxiv.org/abs/2306.00863

[9] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "Istvt: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023.

[10] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "Mre-net: Multi-rate excitation network for deepfake video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3663–3676, 2023.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024. [Online]. Available: https://arxiv.org/abs/2312.00752

[12] X. Xuan, Z. Zhu, W. Zhang, Y.-C. Lin, and T. Kinnunen, "Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative," 2025.

[13] Y. Chen, J. Yi, J. Xue, C. Wang, X. Zhang, S. Dong, S. Zeng, J. Tao, L. Zhao, and C. Fan, "Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection," 2024. [Online]. Available: https://arxiv.org/abs/2406.06086

[14] S. Peng, T. Zhang, L. Gao, X. Zhu, H. Zhang, K. Pang, and Z. Lei, "Wmamba: Wavelet-based mamba for face forgery detection," 2025. [Online]. Available: https://arxiv.org/abs/2501.09617

[15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. [Online]. Available: https://arxiv.org/abs/1905.02244

[16] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," 2023. [Online]. Available: https://arxiv.org/abs/2301.01767