XLSR-MamBo: Scaling the Hybrid Mamba-Attention Backbone for Audio Deepfake Detection

Kwok-Ho Ng[1], Tingting Song[1]*, Yongdong Wu[1], Zhihua Xia[1],

[1]College of Cyber Security, Jinan University, Guangzhou, China
kwokhong@stu2024.jnu.edu.cn, wuyd175@gmail.com, xia_zhihua@163.com
*Correspondence: tingtingsong@jnu.edu.cn

Abstract

Advanced speech synthesis technologies have enabled highly realistic speech generation, posing security risks that motivate research into audio deepfake detection (ADD). While state space models (SSMs) offer linear complexity, pure causal SSMs architectures often struggle with the content-based retrieval required to capture global frequency-domain artifacts. To address this, we explore the scaling properties of hybrid architectures by proposing XLSR-MamBo, a modular framework integrating an XLSR front-end with synergistic Mamba-Attention backbones. We systematically evaluate four topological designs using advanced SSM variants, Mamba, Mamba2, Hydra, and Gated DeltaNet. Experimental results demonstrate that the MamBo-3-Hydra-N3 configuration achieves competitive performance compared to other state-of-the-art systems on the ASVspoof 2021 LA, DF, and In-the-Wild benchmarks. This performance benefits from Hydra's native bidirectional modeling, which captures holistic temporal dependencies more efficiently than the heuristic dual-branch strategies employed in prior works. Furthermore, evaluations on the DFADD dataset demonstrate robust generalization to unseen diffusion- and flow-matching-based synthesis methods. Crucially, our analysis reveals that scaling backbone depth effectively mitigates the performance variance and instability observed in shallower models. These results demonstrate the hybrid framework's ability to capture artifacts in spoofed speech signals, providing an effective method for ADD. Codes are publicly available at https://github.com/saki-ciallo/XLSR-MamBo.

1 Introduction

Driven by rapid advances in generative artificial intelligence (GenAI), cutting-edge text-to-speech (TTS) and voice conversion (VC) technologies have achieved remarkable progress (Casanova et al., 2024; Du et al., 2025; Zhou et al., 2025). The naturalness and timbre similarity of synthesized speech have reached levels that are increasingly challenging for the human ear to distinguish. While beneficial for applications like audiobooks and content creation, these technologies pose severe risks if misused (Li et al., 2025). The maturation of zero-shot and few-shot voice cloning allows attackers to synthesize highly realistic forged speech using only short audio segments collected from social media (Azzuni and Saddik, 2025). Such misuse not only facilitates the spread of misinformation and erodes societal trust but also threatens voice biometric systems (Barrington et al., 2025; McConvey, 2025). Consequently, developing robust and effective audio deepfake detection (ADD) models is imperative.

Early research in anti-spoofing primarily relied on handcrafted features to detect synthesis artifacts that deviate from human speech production mechanisms (Khan et al., 2023; Yang et al., 2026). Although such methods offer interpretability and computational efficiency, these often lack generalization against unseen advanced generators. With the advancement of deep learning, mainstream ADD models have shifted to end-to-end architectures, which directly process raw audio for feature extraction and classification. Regarding front-end feature extraction, XLSR (Babu et al., 2022) is a self-supervised learning (SSL) model based on Wav2Vec 2.0 (Baevski et al., 2020), pre-trained on massive cross-lingual datasets. Benefiting from this large-scale diversity, XLSR has demonstrated competitive generalization capabilities compared to other pre-trained models (PTMs) such as Whisper (Radford et al., 2023) and WavLM (Chen et al., 2022), demonstrating strong competitiveness in ADD tasks (Kashyap et al., 2025). Building on this powerful representation, attention-based classifiers have become an efficient back-end candidate. The Conformer architecture (Gulati et al., 2020), which

Findings of the Association for Computational Linguistics: ACL 2026, pages 31450–31462

combines the global context modeling of Transformers with the local feature capture of CNNs, is widely adopted. By integrating XLSR with Conformer, the XLSR-Conformer (Rosello et al., 2023) achieved state-of-the-art (SOTA) performance on the ASVspoof 2021 LA (ASV21LA) and DF (ASV21DF) (Yamagishi et al., 2021) dataset at the time. Subsequent variants, such as XLSR-Conformer+TCM Truong et al. (2024), XLSR-SLS Zhang et al. (2024), and Kanformer (Dat and Dat, 2025), further pushed the performance boundaries by refining the attention mechanism.

However, the quadratic computational complexity of the classical Transformer limits its efficiency in modeling dense feature sequences. To address this, state space models (SSMs) (Gu et al., 2022) with linear complexity, particularly the Mamba architecture (Gu and Dao, 2024), have attracted widespread attention. RawBMamba (Chen et al., 2024) was among the early works to explore the application of Mamba in audio deepfake detection. Addressing the unidirectional limitations of the original Mamba, it proposed a Bidirectional Mamba architecture that combines SincLayers/convolutional layers (to capture short-term local features) with bidirectional Mamba blocks (to capture long-range contextual dependencies). This design effectively identifies subtle artifacts in synthesized audio and significantly outperforms traditional end-to-end models like Rawformer (Liu et al., 2023), validating the generalization capability of the Mamba architecture in audio forensics. Following this, works like XLSR-Mamba (Xiao and Das, 2025) and Fake-Mamba (Xuan et al., 2025) further advanced the field. The end-to-end BiCrossMamba-ST (Kheir et al., 2025) integrates BiMamba with cross-attention, achieving better performance than RawBMamba. However, a critical observation in these works is their reliance on manually designed bidirectional strategies (e.g., dual-branch fusion) to approximate the non-causal receptivity of Transformers. While these studies successfully validated that substituting Transformer layers with SSM blocks yields efficiency gains, relying solely on this replacement strategy overlooks the unique complementary strengths of different mechanisms.

Although pure SSM models excel in inference efficiency and ADD performance, experimental results from cutting-edge hybrid architectures such as Zamba (Glorioso et al., 2024), Samba (Ren et al., 2025), Jamba (Lieber et al., 2025), and Nemotron-H (Blakeman et al., 2025) demonstrate that combining SSMs with Attention yields competitive performance. Furthermore, hardware-aware algorithms such as FlashAttention (Dao et al., 2022) and Mesh-Attention (Chen et al., 2025) have alleviated the computational bottlenecks of Transformers. From the perspective of inductive bias, SSMs specialize in information compression and temporal recurrence, whereas Attention excels at content-based, precise retrieval and associative recall via the induction heads mechanism (Olsson et al., 2022; Arora et al., 2024). Consequently, in ADD tasks, the Attention mechanism can be better suited to globally correlate specific frequency-domain forgery artifacts. Specifically, deepfake signals often manifest as a duality: subtle local high-frequency artifacts (requiring the fine-grained temporal recurrence of SSMs) and global spectral inconsistencies (requiring the global retrieval of Attention). Relying solely on one mechanism limits the model's ability to capture this full spectrum of manipulation traces (Frank et al., 2020; Sun et al., 2023).

In light of this, we propose a novel framework, XLSR-MamBo, for anti-spoofing. This framework employs XLSR as the front-end and utilizes a hybrid SSM-Attention architecture as the back-end encoder. Compared to prior works that rely on manual bidirectionality, we explore advanced variants including Mamba2 (Dao and Gu, 2024), Gated DeltaNet (GDN) (Yang et al., 2024), and notably, Hydra (Hwang et al., 2024). Hydra introduces a paradigm shift by parameterizing the sequence mixer as a quasiseparable matrix. Unlike heuristic dual-branch extensions, this formulation allows Hydra to perform native bidirectional processing with linear complexity, capturing holistic non-causal dependencies without structural redundancy. This theoretical advantage positions it as an ideal candidate for detecting artifact traces that violate causal consistency. Furthermore, we investigate the impact of scaling the backbone depth to mitigate inference instability. The proposed framework is evaluated on the ASV21LA and DF, In-the-Wild (ITW) (Müller et al., 2022), and DFADD (Du et al., 2024) datasets, and the results demonstrate its robustness and effectiveness compared to other SOTA systems.

2 Background

State Space Model. The basic SSM defines a continuous-time linear time-invariant (LTI) system. It maps a 1-dimensional input signal $x(t) \in \mathbb{R}$ to

a latent state $h(t) \in \mathbb{R}^N$, and then projects it to an output $y(t) \in \mathbb{R}$:

$$h'(t) = Ah(t) + Bx(t),\ y(t) = Ch(t) \quad (1)$$

where $A \in \mathbb{R}^{N\times N}$ is the state transition matrix, $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{1\times N}$ are the input and output projection matrices.

To process discrete sequence data, the continuous system is discretized using a learnable step size Δ, representing the sampling interval. Applying the zero-order hold (ZOH) rule yields the discrete parameters:

$$\begin{aligned} \bar{A} &= \exp(\Delta A) \\ \bar{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \end{aligned} \quad (2)$$

The discretized SSM state update equation is defined as follows:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t,\ y_t = Ch_t \quad (3)$$

Selective SSM, Mamba. Because the basic LTI SSM uses fixed $\bar{A}$ and $\bar{B}$ matrices across all time steps, it cannot dynamically adjust its memory strategy. Mamba solves this by introducing a selective mechanism where parameters map dynamically from the input, i.e., $x_t \to (\Delta_t, B_t, C_t)$:

$$\begin{aligned} \Delta_t &= \text{Softplus}(\text{Linear}_\Delta(x_t)) \\ B_t &= \text{Linear}_B(x_t),\ C_t = \text{Linear}_C(x_t) \\ \bar{A}_t &= \exp(\Delta_t A),\ \bar{B}_t = \Delta_t B_t \end{aligned} \quad (4)$$

The Mamba state update equation becomes:

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t,\ y_t = C_t h_t \quad (5)$$

State Space Duality (SSD), Mamba2. While Mamba's dynamic parameterization enhances expressivity, it reduces computational efficiency. Mamba2 introduces the SSD framework, structurally constraining the continuous state matrix A (e.g., to a scalar multiple of the identity matrix). This simplifies the discretization process and allows the state update to be reformulated as a decay-modulated linear attention, represented by the compact recurrence:

$$S_t = \alpha_t S_{t-1} + k_t v_t^\top,\ y_t = q_t^\top S_t \quad (6)$$

where $\alpha_t \in (0, 1)$ is a input -dependent decay factor, and S_t is the hidden state matrix at time step t. The mapping $x_t \to (q_t, k_t, v_t)$ generates the query, key, and value vectors corresponding to the attention mechanism. This formulation enables a chunkwise parallel algorithm, significantly accelerating computation via matrix multiplication.

Hydra. Standard SSMs are causal recurrent systems. To adapt them for non-causal tasks, Hydra parameterizes the combination of forward and backward scans as a quasiseparable matrix. This mixed matrix contains both lower-triangular (past information propagation) and upper-triangular (future information propagation) semiseparable structures. The Hydra's formula is defined as follow:

$$\text{shift}(SS(X)) + \text{flip}(\text{shift}(SS(\text{flip}(X)))) + DX \quad (7)$$

where X is the input sequence, $\text{SS}(\cdot)$ denotes the SSD, $\text{shift}(\cdot)$ and $\text{flip}(\cdot)$ denotes a right-shift and reverse function, and $D = \text{diag}(\delta_1, \ldots, \delta_L)$ represents the diagonal elements of the matrix.

Gated DeltaNet. To overcome the memory capacity limits of linear attention, GDN integrates the delta rule with Mamba2's dynamic gating decay. This mechanism actively erases components of the prior state that conflict with new information (i.e., those aligned with the current key k_t). The GDN recurrence is defined as follows:

$$S_t = \alpha_t(I - \beta_t k_t k_t^\top)S_{t-1} + \beta_t k_t v_t^\top \quad (8)$$

where β_t denotes the input-dependent write strength. The term $S_{t-1} \to S_{t-1}(I - \beta_t k_t k_t^\top)$ acts as a soft-elimination operator, projecting S_{t-1} into a subspace orthogonal to k_t to minimize interference. For efficiency, the WY representation is used to parallelize the recurrence into chunks, accelerating the training.

3 Methods

The original Mamba model performs causal computations in a unidirectional manner, utilizing only historical information. However, audio deepfake detection necessitates processing complete utterances to capture global dependencies and spectral inconsistencies. While bidirectional adaptations of SSMs can address contextuality, recent advancements in large language models (LLMs) such as Jamba suggest that hybridizing SSMs with Attention mechanisms offers a superior inductive bias compared to pure SSM architectures. This hybrid paradigm effectively combines the efficient temporal compression of SSMs with the precise content-based retrieval of Attention. Yet, the efficacy of such synergistic designs remains under-

explored within the specific domain of audio anti-spoofing. To systematically assess the potential of this hybrid synergy, we propose a unified modular framework—XLSR-MamBo. Unlike previous research that primarily focuses on pure SSM backbones, this work is dedicated to exploring diverse topological combinations of SSM variants within hybrid SSM-Attention architectures.

The hybrid design is motivated by the complementary strengths of SSMs (efficient sequence modeling) and Attention mechanisms (precise global context recall). Inspired by MambaFormer (Park et al., 2024), we place the Mamba block after the input projection layer. This configuration leverages the implicit positional encoding inherent in the recurrent structure of SSMs, mitigating the need for explicit positional encodings while maintaining robust sequence modeling (Gu and Dao, 2024). All blocks incorporate Pre-Norm (Xiong et al., 2020) connections to enhance training stability.

As shown in Figure 1, we explore four distinct layer architectures. Each structurally resembles a standard Transformer but modifies its internal modules to analyze different syntheses of SSMs and Attention.

MamBo-1 (Mamba Layer, Figure 1 (a)): Replaces the multi-head self-attention (MHA) with a Mamba block while retaining the subsequent SwiGLU (Shazeer, 2020) feed-forward network (FFN).

$$H' = [H + \text{Mamba}(\text{Norm}(H))]^{\times N} \quad (9)$$

$$\hat{H} = H' + \text{FFN}(\text{Norm}(H')) \quad (10)$$

This provides a baseline to isolate the standalone effectiveness of SSM variants in capturing long-range dependencies without explicit attention.

MamBo-2 (Mamer Layer, Figure 1 (b)): Replaces the FFN in the Mamba Layer with a MHA block. This represents a compact intra-block hybrid, placing SSM and Attention in immediate succession.

$$\hat{H} = H' + \text{MHA}(\text{Norm}(H')) \quad (11)$$

This design aims to balance efficient sequence compression with precise global associations within a single block unit.

MamBo-3 (Mamba-Transformer Layer, Figure 1 (c)): Adopts a classic interleaved strategy by alternating standard Mamba Layers and Transformer Layers. Building on recent works (Zhang et al., 2025) that position Mamba as a promising alternative to self-attention in speech modeling. This sequential hybridization provides a baseline for evaluating the benefits of separating SSM processing and Attention-based refinement into distinct layers.

MamBo-4 (Mamba-Mamer Layer, Figure 1 (d)): Further intensifies the hybrid mechanism by alternating Mamba Layers and Mamer Layers. This high-density hybrid explores whether a higher concentration of SSM-Attention interactions yields improved detection of complex spoofing artifacts.

Additionally, we introduce a stacking hyperparameter N, inspired by Zamba. This configuration allows for the stacking of N consecutive SSM blocks within a single unit, facilitating a systematic evaluation of depth scaling in SSM modules and its impact on detection performance in audio deepfake detection tasks.

As shown in Figure 1, the overall pipeline of the proposed XLSR-MamBo is structured as follows. The input raw waveform is first processed by the pre-trained XLSR model, which extracts a sequence of high-level speech representations, resulting in a feature matrix $X \in \mathbb{R}^{T\times 1024}$, where T denotes the number of time frames proportional to the input audio duration. To stabilize training and align feature dimensionality, we apply RMSNorm (Zhang and Sennrich, 2019) followed by a linear projection layer, mapping the channel dimension from 1024 to the hidden dimension D. The resulting projected features $H \in \mathbb{R}^{T\times D}$ serve as input to the backbone encoder, which consists of stacked MamBo layers (L).

$$H = \text{Linear}_{\text{down}}(X) \quad (12)$$

$$\hat{H} = \text{MamBo}(H)^{\times L} \in \mathbb{R}^{T\times D} \quad (13)$$

This encoder effectively captures complex temporal dependencies and subtle manipulation artifacts while preserving the feature shape $(T \times D)$. Subsequently, the frame-level features are aggregated into a fixed-dimensional utterance-level representation using gated attention pooling. Finally, a single linear layer projects this pooled representation to a 2-dimensional output, yielding logits for binary classification (spoof vs. bonafide).

4 Experimental Setup

4.1 Datasets and Metrics

All MamBo variants were trained on the ASVspoof 2019 LA (ASV19LA) training set (Wang et al.,

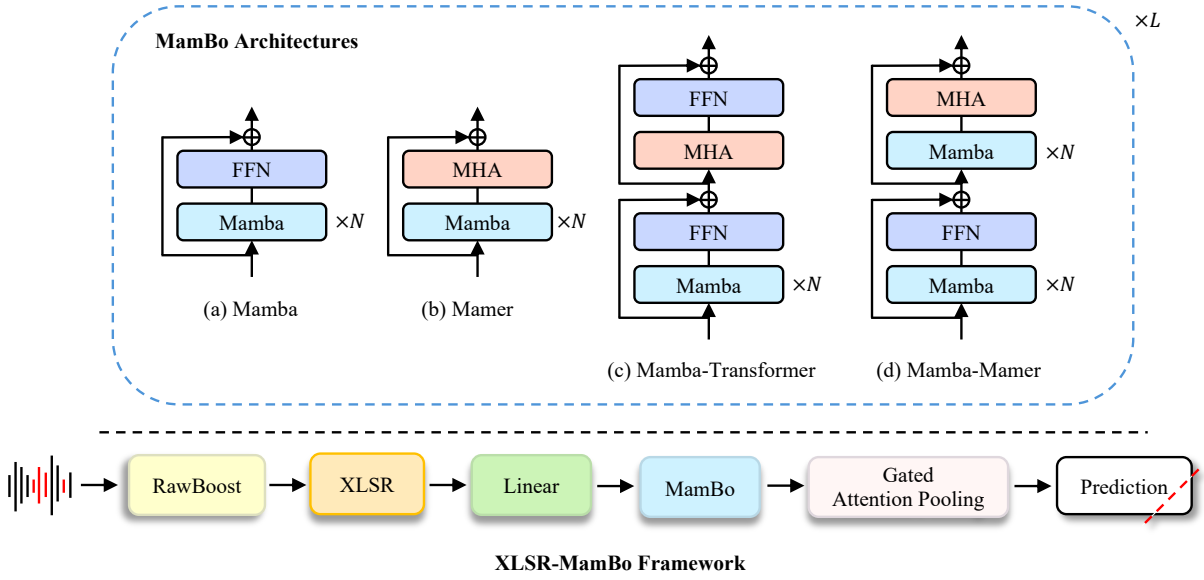

Figure 1: The overall proposed XLSR-MamBo framework. The Mamba block is depicted as the representative instantiation of the SSM component. Four variant configurations of the MamBo architectures include: (a) replacing the MHA in Transformer layer with an SSM module, termed Mamba; (b) substituting the FFN in a Mamba layer with MHA to capture non-causal dependencies, termed Mamer; (c) combining Mamba layer with Transformer layers (non-causal and without positional encoding), termed Mamba-Transformer; and (d) hybridizing Mamba layer and Mamer layer, termed Mamba-Mamer.

2020), which comprises 25,380 training, 24,844 development, and 71,237 evaluation utterances. To evaluate the generalizability and robustness of the proposed method via cross-dataset evaluation, we conducted experiments on four challenging test sets: ASV21LA (181,566 utterances), ASV21DF (611,829 utterances), ITW (31,778 utterances), and DFADD (1,355 utterances each subset). The ASVspoof 2021 introduces complex transmission channel variations and compression artifacts. The ITW dataset contains real-world deepfake audio sourced from social media, while DFADD includes synthetic speech generated by advanced diffusion- and flow-matching-based TTS models (subsets D1–D3 and F1–F2, respectively). Following the standard protocols of the ASVspoof Challenges, we report the equal error rate (EER) and the minimum tandem detection cost function (min t-DCF) as evaluation metrics.

4.2 Implementation Details

For data augmentation, we used RawBoost (Tak et al., 2022). Consistent with baseline configurations (Xiao and Das, 2025; Xuan et al., 2025), we trained separate models for the ASV21LA (21LA) and ASV21DF (21DF) evaluations. The 21LA model applied linear and non-linear convolutive noise combined with impulsive signal-dependent noise, while the 21DF model used stationary signal-independent noise.

All audio inputs were processed at 16 kHz with a fixed duration of 4.175s (66,800 samples). We fine-tuned the models using the AdamW (Loshchilov and Hutter, 2019) optimizer ($lr = 10^{-5}$, $\beta_1 = 0.9, \beta_2 = 0.95$, weight decay=0.05) with a scheduler combining 10% linear warmup with cosine decay. Training was performed in mixed precision (BF16/FP32) using FocalLoss (Lin et al., 2017) to mitigate class imbalance, for a maximum of 20 epochs with early stopping (patience of 7 epochs), and a batch size of 32. The top-5 checkpoints with the lowest validation loss were retained. For the architecture, the projection dimension was set to $D = 128$, with $L = 5$ MamBo layers by default. For the Mamba variants (Mamba2, Hydra, GDN), the state dimension (d_state) and head dimension were set to 64 and 32, respectively. All experiments were conducted on one RTX 4080 Super 16 GB and two RTX 3090 24 GB GPUs. To facilitate reproducibility, a fixed random seed was used in experimental runs.

5 Results and Analysis

5.1 MamBo-1 and MamBo-2

Tables 1 and 2 summarize the performance of the MamBo series architectures across SSM variants (Mamba, Mamba-2, Hydra, GDN) and SSM stacking depths N. To evaluate detection capability against advanced generative algorithms, results on the DFADD dataset are reported separately for the D1-D3 subsets (generated by distinct diffusion models) and F1-F2 subsets (generated by distinct flow-matching models).

In the evaluation of the MamBo-1 architecture on ASV21LA, Mamba, Mamba-2, and Hydra achieved their lowest EERs (0.81%, 0.79%, and 0.92%, respectively) and best min t-DCF scores at a stacking depth of $N = 2$. On ASV21DF, however, the variants showed divergent depth preferences: Mamba and Mamba-2 maintained competitive performance at $N = 2$, whereas Hydra and GDN demonstrated advantages with a shallower configuration of $N = 1$. This divergence became more evident on the ITW dataset, where Mamba and Mamba-2 demonstrated stronger generalization at $N = 1$, while Hydra and GDN required $N = 3$ to achieve the best generalization performance. These data patterns indicate that, within the same architectural framework, different SSM variants exhibit distinct inductive biases, resulting in varying dependencies on stacking depth for generalization to out-of-domain datasets.

Regarding the more challenging DFADD dataset, which presents substantially greater challenges than in-domain datasets, all variants produced comparable and consistently low EERs on D1-D3, demonstrating the model's efficacy in robustly detecting of diffusion-based spoofing attacks. On the more difficult F1-F2 subsets, performance of Mamba, Mamba-2, and Hydra varied non-monotonically with stacking depth: strong generalization at $N = 1$ and $N = 3$, but degradation at $N = 2$, suggesting potential optimization challenges at intermediate depths. In contrast, GDN maintained a shallow-layer advantage, achieving its best EERs at $N = 1$ (5.83%) and $N = 2$ (3.17%). The overall data trend suggests that increasing SSM stacking depth generally contributed to improved generalization on complex out-of-domain data, notwithstanding variant-specific fluctuations.

For the MamBo-2 architecture, performance remained relatively balanced across configurations on ASV21LA and DF. On the ITW dataset, however, deeper stacks consistently lowered EERs, with Hydra at $N = 3$ attaining the minimum value (3.80%), representing a relative improvement of approximately 39% over $N = 1$. Similar scaling benefits emerged on the challenging F1 subset, where Mamba ($N = 3$) and Mamba-2 ($N = 3$) achieved substantial EER reductions to 3.84% (approx. 65% relative improvement) and corresponding gains of approx. 71%, respectively. The GDN maintained competitive performance in shallower configurations ($N = 1$), consistent with observations in MamBo-1. These findings suggest that denser SSM-Attention hybridization amplifies scaling benefits with increased stacking depth, enabling more effective identification of subtle artifacts from unknown generative models.

5.2 MamBo-3 and MamBo-4

Table 2 further presents results for the more complex hybrid architectures, MamBo-3 and MamBo-4. On the ASV21LA (MamBo-3), Mamba, Mamba-2, and Hydra attained their lowest EERs (0.96%, 0.94%, and 0.81%, respectively) at $N = 3$, with min t-DCF scores generally correlating positively with EER. However, GDN optimized EER (0.96%) at $N = 1$ but achieved the best min t-DCF at $N = 2$. This difference suggests that while the shallower configuration finds a better equilibrium point for the EER, increasing the depth to $N = 2$ enhances score calibration, making it more robust under the weighted cost constraints prioritized by the min t-DCF metric.

On the ASV21DF, performance trends diverged across variants: Mamba and Mamba-2 performed best with shallower stacks ($N = 2$ for Mamba at 1.77% EER; $N = 1$ for Mamba-2 at 1.56% EER), effectively addressing compression artifacts. Hydra and GDN, conversely, attained lowest EERs (1.70% and 1.82%, respectively) at $N = 3$, suggesting that greater depth improves robustness to channel variations. Generalization on ITW exhibited analogous patterns: Mamba achieved its lowest EER (4.78%) at $N = 2$, Mamba-2 and Hydra at $N = 3$ (4.45% and 4.97%, respectively). These results show that increasing model capacity contributes to the effective capture of out-of-domain features.In contrast, GDN maintained its dominance in the shallow layers, achieving a competitive EER of 4.72% at $N = 1$.

In the detailed evaluation on the DFADD dataset, results on D1-D3 aligned closely with ITW trends. The Mamba ($N = 2$) and Mamba-2 ($N = 3$) per-

Model	N	ASV21LA		ASV21DF	ITW	D1	D2	D3	F1	F2
		min t-DCF ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓
MamBo-1										
Mamba	1	0.2178 / 0.2196	1.19 / 1.27	2.08 / 2.88	**4.65** / 4.66	2.87 / 3.28	5.17 / 5.68	0.15 / 0.31	11.36 / 14.64	24.35 / 26.47
	2	**0.2060** / 0.2136	**0.81** / 1.10	**2.00** / 5.03	5.88 / 5.96	4.14 / 4.18	**1.18** / 1.30	0.51 / 0.69	30.18 / 31.83	**20.81** / 22.95
	3	0.2139 / 0.2217	1.14 / 1.43	2.15 / 2.47	5.51 / 5.80	**1.33** / 1.42	1.48 / 1.48	0.00 / 0.19	**11.00** / 14.56	21.69 / 22.83
Mamba2	1	0.2170 / 0.2209	1.21 / 1.34	2.47 / 5.13	**4.04** / 5.10	**0.51** / 0.80	**0.51** / 1.45	0.00 / 0.00	**7.16** / 18.79	16.46 / 20.35
	2	**0.2056** / 0.2113	**0.79** / 0.96	**2.01** / 2.41	5.57 / 8.14	1.69 / 6.43	1.69 / 4.38	0.00 / 0.03	9.00 / 18.08	**12.54** / 23.51
	3	0.2108 / 0.2154	0.98 / 1.13	2.37 / 3.19	5.27 / 6.59	1.99 / 2.34	1.84 / 2.50	0.00 / 0.13	7.46 / 14.25	18.82 / 21.88
Hydra	1	0.2161 / 0.2181	1.11 / 1.19	**2.06** / 2.74	6.91 / 6.99	**0.81** / 0.95	**0.81** / 0.98	0.15 / 0.15	6.20 / 7.33	**12.48** / 13.37
	2	**0.2100** / 0.2140	**0.92** / 1.06	2.10 / 3.17	7.18 / 7.96	3.02 / 8.61	1.03 / 5.89	0.00 / 2.43	16.83 / 31.00	16.46 / 39.47
	3	0.2109 / 0.2300	0.97 / 1.69	2.10 / 2.44	**5.22** / 5.33	1.84 / 2.03	1.69 / 2.00	0.00 / 0.00	**4.20** / 6.46	13.36 / 16.44
GDN	1	0.2118 / 0.2175	1.01 / 1.22	**2.12** / 2.30	5.98 / 6.08	0.81 / 0.95	0.96 / 1.06	0.00 / 0.00	5.83 / 6.65	14.53 / 15.15
	2	0.2231 / 0.2349	1.38 / 1.84	2.12 / 2.71	6.29 / 6.92	**0.66** / 0.82	**0.96** / 1.04	0.37 / 0.87	**3.17** / 6.59	**9.30** / 12.68
	3	**0.2110** / 0.2140	**0.95** / 1.08	2.68 / 3.75	**4.16** / 4.17	3.69 / 3.72	1.03 / 1.15	0.00 / 0.00	15.65 / 16.42	20.37 / 21.13
MamBo-2										
Mamba	1	0.2102 / 0.2145	0.95 / 1.11	2.54 / 2.99	6.34 / 8.78	**1.03** / 2.70	1.48 / 3.38	0.00 / 0.09	11.15 / 16.72	21.18 / 26.31
	2	**0.2056** / 0.2073	**0.79** / 0.83	2.19 / 2.29	**4.93** / 6.86	2.14 / 4.09	1.33 / 3.09	0.00 / 0.00	7.97 / 12.24	14.53 / 16.78
	3	0.2138 / 0.2172	1.05 / 1.18	**2.09** / 2.27	6.15 / 6.57	1.69 / 2.09	**0.81** / 2.21	0.00 / 0.00	**3.84** / 4.40	**14.47** / 15.83
Mamba2	1	0.2132 / 0.2216	1.06 / 1.37	**2.02** / 3.35	7.23 / 7.32	1.63 / 1.66	1.84 / 2.11	0.00 / 0.00	22.14 / 24.39	23.02 / 25.19
	2	**0.2092** / 0.2237	**0.90** / 1.41	2.11 / 2.62	7.53 / 7.62	**0.81** / 1.04	**0.81** / 0.91	0.15 / 0.15	8.34 / 8.88	**11.36** / 12.53
	3	0.2117 / 0.2183	1.00 / 1.21	2.39 / 3.14	**4.09** / 4.12	1.48 / 1.59	1.03 / 1.12	0.15 / 0.15	**6.35** / 6.96	16.68 / 16.97
Hydra	1	**0.2066** / 0.2199	**0.80** / 1.32	**1.84** / 2.73	6.24 / 6.56	1.84 / 2.72	1.69 / 2.37	0.00 / 0.03	**5.32** / 8.54	8.85 / 23.93
	2	0.2116 / 0.2134	1.00 / 1.05	2.28 / 2.94	5.01 / 5.12	1.99 / 2.59	1.69 / 2.56	0.00 / 0.00	10.33 / 13.29	20.30 / 22.71
	3	0.2239 / 0.2257	1.43 / 1.48	2.08 / 2.34	**3.80** / 5.22	**0.37** / 1.28	**0.37** / 1.84	0.00 / 0.00	5.98 / 12.56	**6.79** / 14.61
GDN	1	0.2185 / 0.2313	1.24 / 1.73	2.30 / 2.38	6.52 / 6.89	1.99 / 2.68	**1.03** / 1.21	0.00 / 0.03	**7.97** / 9.94	**17.19** / 21.38
	2	**0.2092** / 0.2175	**0.89** / 1.18	1.96 / 2.45	**5.27** / 6.59	5.68 / 6.44	7.97 / 8.82	0.00 / 0.00	24.35 / 25.20	29.97 / 30.83
	3	0.2124 / 0.2156	1.01 / 1.11	**1.94** / 2.63	7.12 / 7.30	**1.99** / 2.16	2.51 / 2.66	0.37 / 0.39	18.16 / 19.92	27.97 / 29.81

Table 1: Comparison of MamBo-1 and MamBo-2 architectures across varying stacking depths N on the ASV21LA, ASV21DF, ITW, and DFADD evaluation sets. All models were trained on ASV19LA. Results are reported as "Best / Avg" across the top-5 checkpoints.

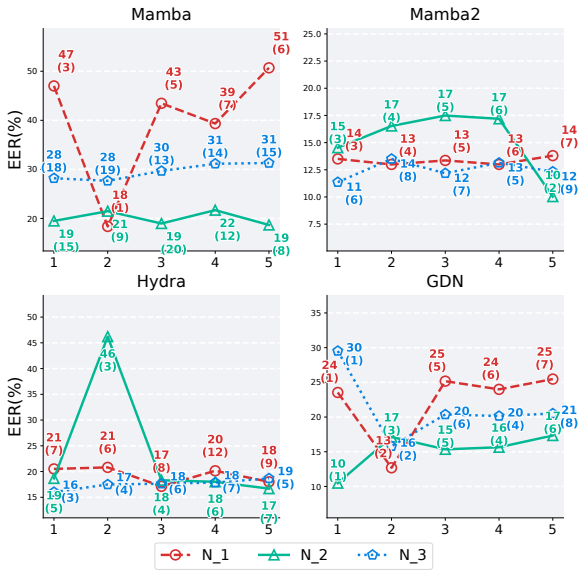

Figure 2: Evaluation results of the MamBo-3 architecture utilizing four distinct SSM variants across varying stacking depths N on the DFADD-F2 subset. The x-axis shows the top-5 checkpoints ranked by validation loss (where 1 indicates the lowest loss). Values in parentheses indicate the corresponding training epoch for each checkpoint.

formed best, GDN remained dominated by shallow settings, and Hydra peaked at $N = 2$. However, on the more challenging F1-F2 subsets, all four variants exhibited improved performance through appropriate stacking compared to shallower settings. The checkpoint EERs trends for the F2 dataset are illustrated in Figure 2. Notably, Mamba-2 achieved the lowest EER of 3.02% on the F1 dataset with the $N = 3$ configuration, while Hydra demonstrated comparable performance of 3.47% at $N = 2$.

The experimental evaluation of the MamBo-4 architecture involves a comparative analysis between the standard configuration ($L = 5$) and a deep ablation setting ($L = 7$). Under the default $L = 5$ configuration, Mamba and Mamba-2 exhibited comparable performance on ASV21LA, whereas GDN attained the minimum EER of 0.86%. On the ASV21DF and ITW datasets, Hydra demonstrated notable efficacy, securing the lowest EERs of 1.43% and 5.17%, respectively. Regarding the DFADD evaluation, the four SSM variants showed negligible performance differences across the D1-D3 subsets. In contrast, on the F1-F2 subsets, Mamba achieved the lowest EERs (5.68% on F1, 7.16% on F2) but exhibited substantial checkpoint

Model	N	ASV21LA		ASV21DF	ITW	D1	D2	D3	F1	F2
		min t-DCF ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓	EER(%) ↓
MamBo-3										
Mamba	1	0.2272 / 0.2333	1.53 / 1.77	1.80 / 2.59	7.54 / 8.27	8.85 / 10.56	1.69 / 2.73	0.00 / 0.06	16.98 / 33.37	**18.37** / 39.78
	2	0.2118 / 0.2180	1.09 / 1.30	**1.77** / 2.68	**4.78** / 4.82	**0.51** / 0.63	**0.81** / 0.91	0.15 / 0.18	**8.34** / 9.03	18.67 / 20.06
	3	**0.2107** / 0.2149	**0.96** / 1.14	2.01 / 2.84	5.71 / 5.73	1.33 / 1.45	4.80 / 5.22	0.15 / 0.15	9.30 / 10.19	27.67 / 29.59
Mamba2	1	0.2196 / 0.2296	1.41 / 1.64	**1.56** / 1.88	6.04 / 6.08	1.33 / 1.49	0.66 / 0.82	0.15 / 0.25	6.20 / 6.61	12.99 / 13.32
	2	0.2193 / 0.2442	1.26 / 2.20	2.15 / 3.04	4.83 / 5.12	1.03 / 1.30	0.66 / 1.13	0.00 / 0.25	7.46 / 10.46	**10.03** / 15.15
	3	**0.2112** / 0.2145	**0.94** / 1.08	2.02 / 2.37	**4.45** / 4.48	**0.15** / 0.15	**0.30** / 0.30	0.00 / 0.00	**3.02** / 3.55	11.36 / 12.50
Hydra	1	0.2222 / 0.2264	1.34 / 1.49	2.58 / 4.79	6.00 / 6.05	2.66 / 2.90	3.84 / 4.04	0.00 / 0.00	14.47 / 15.70	17.13 / 19.32
	2	0.2152 / 0.2308	1.11 / 1.68	2.12 / 2.98	5.58 / 6.43	**0.66** / 1.56	**1.18** / 2.68	0.00 / 0.47	**3.47** / 8.44	16.68 / 23.57
	3	**0.2072** / 0.2087	**0.81** / 0.87	**1.70** / 1.81	**4.97** / 5.18	1.84 / 2.53	1.33 / 1.69	0.00 / 0.00	11.36 / 12.73	**16.01** / 17.53
GDN	1	0.2111 / 0.2181	**0.96** / 1.23	2.20 / 3.29	**4.72** / 5.92	**1.03** / 3.53	1.48 / 3.30	0.15 / 0.18	11.51 / 17.59	12.69 / 22.17
	2	**0.2107** / 0.2151	0.99 / 1.23	2.30 / 3.15	5.54 / 6.15	4.20 / 6.03	**1.03** / 1.21	0.00 / 0.00	13.21 / 15.18	**10.48** / 15.19
	3	0.2127 / 0.2152	1.00 / 1.13	**1.82** / 1.99	5.30 / 6.64	1.48 / 2.16	1.03 / 2.56	0.30 / 0.54	**10.18** / 18.14	15.80 / 21.27
MamBo-4										
Mamba		0.2100 / 0.2118	0.90 / 1.00	2.14 / 2.98	5.54 / 5.98	**1.03** / 2.14	1.03 / 1.18	0.15 / 0.29	**5.68** / 20.85	**7.16** / 20.72
Mamba2		0.2093 / 0.2140	0.90 / 1.06	2.17 / 2.50	6.42 / 7.01	3.32 / 4.65	3.69 / 5.00	0.15 / 0.34	16.01 / 23.14	26.79 / 32.85
Hydra		0.2119 / 0.2330	0.98 / 1.74	**1.43** / 2.49	**5.17** / 5.39	1.33 / 2.70	1.84 / 4.66	0.00 / 0.03	14.17 / 24.19	19.34 / 29.30
GDN		**0.2090** / 0.2145	**0.86** / 1.09	2.36 / 2.90	5.42 / 5.72	1.69 / 1.97	**0.66** / 1.12	0.00 / 0.03	7.97 / 17.25	17.13 / 22.59
MamBo-4 with $L = 7$										
Mamba		**0.2084** / 0.2162	0.93 / 1.18	2.52 / 2.65	**4.92** / 5.85	1.99 / 2.70	2.81 / 3.61	0.15 / 0.44	14.47 / 16.53	23.99 / 25.69
Mamba2		0.2226 / 0.2291	1.38 / 1.62	1.94 / 2.44	5.26 / 6.21	1.03 / 1.42	**0.37** / 0.51	0.00 / 0.23	8.49 / 9.02	11.36 / 12.38
Hydra		0.2090 / 0.2123	**0.88** / 0.99	**1.76** / 2.45	5.60 / 5.63	**0.51** / 0.75	1.18 / 1.27	0.00 / 0.00	7.82 / 9.79	12.69 / 15.49
GDN		0.2142 / 0.2244	1.11 / 1.45	1.94 / 2.28	6.05 / 6.29	0.66 / 1.25	0.37 / 1.81	0.00 / 0.36	**5.98** / 8.32	**9.00** / 18.70

Table 2: Comparison of the MamBo-3 architecture across varying stacking depths N and MamBo-4 under different backbone depths ($L = 5$ vs. $L = 7$, with fixed $N = 1$) on the ASV21LA, ASV21DF, ITW, and DFADD evaluation sets. All models were trained on ASV19LA. Results are reported as "Best / Avg" across the top-5 checkpoints.

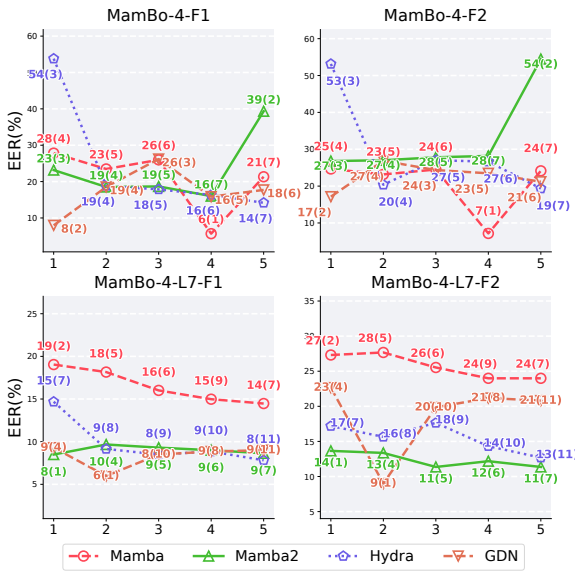

Figure 3: Evaluation results of the MamBo-4 architecture integrating four distinct SSM variants on the DFADD-F1 and F2 subsets. The top and bottom rows display performance with backbone depths $L = 5$ and $L = 7$, respectively (with fixed stacking depth N). The x-axis shows the top-5 checkpoints ranked by validation loss (where 1 indicates the lowest loss). Values in parentheses indicate the corresponding training epoch for each checkpoint.

instability, with only the best of the top-5 checkpoints achieving single-digit EERs, while the remaining checkpoints showed significant deviation, as shown in Figure 3.

In the ablation with increased backbone depth ($L = 7$), deeper architectures generally improved generalization and inference stability across variants on F1-F2 datasets. Notably, Mamba-2 and Hydra demonstrated markedly improved generalization on the F1 dataset, achieving EERs of 8.49% (approx. 46% improvement) and 7.82% (approx. 44% improvement), respectively. However, divergent trade-offs emerged; Mamba performance peaked at $L = 5$ and degraded at $L = 7$, implying potential overfitting despite improved stability, whereas GDN exhibited enhanced generalization but remained the least stable. Overall, Mamba-2 and Hydra demonstrated better cross-domain robustness, with Hydra securing the lowest EERs on ASV21LA (0.88%) and DF (1.76%), confirming the efficacy of increasing backbone depth as an optimization strategy.

A comprehensive analysis of the top-5 checkpoint performance reveals persistent low bias accompanied by high variance. Specifically, while the top-1 checkpoint achieved minimal validation loss on ASV19LA and improved generalization

on DFADD, substantial inference disparity was observed across the remaining candidates. This instability indicates that the decision boundaries of shallow architectures remain sensitive to unseen generative algorithms. Consistent findings from the MamBo-3 and MamBo-4 experiments on F1-F2 demonstrate that increasing model depth mitigates this issue to varying degrees. Consequently, deeper configurations enhance inter-checkpoint consistency and effectively attenuate the performance variance prevalent in shallower models.

5.3 Compare with Other SOTA Systems

Method	Params (M)	21LA	21DF	ITW
		EER	EER	EER
RawMamba[1]	0.719	2.84	22.48	-
RawBMamba[1]	0.719	3.28	15.85	-
BiCrossMamba-ST[2]	0.516	3.39	14.77	-
XLSR-Conformer[3]	319.74	1.38	2.27	-
XLSR-Conformer+TCM[4]	319.77	1.03	2.06	7.79
XLSR-SLS[5]	341.49	2.87	1.92	7.46
XLSR-Mamba[6]	319.33	0.93	1.88	6.71
Fake-Mamba (L)[7]	319.72	0.97	1.74	5.85
MamBo-1-Mamba2-N2	317.48	**0.79**	2.01	5.57
MamBo-2-Hydra-N1	316.68	0.80	1.84	6.24
MamBo-3-Hydra-N3	319.37	0.81	1.70	**4.97**
MamBo-4-Hydra-N1	318.02	0.98	**1.43**	5.17

[1]Chen et al. (2024); [2]Kheir et al. (2025); [3]Rosello et al. (2023); [4]Truong et al. (2024); [5]Zhang et al. (2024); [6]Xiao and Das (2025); [7]Xuan et al. (2025).

Table 3: Comparison of selected configurations from each MamBo variant with SOTA single systems on the 21LA, 21DF, and ITW evaluation sets. Included lightweight end-to-end models that did not use PTM. All baselines trained on ASV19LA.

Table 3 presents the performance of the proposed MamBo configurations (with backbone depth fixed at $L = 5$) in comparison with state-of-the-art models. Experimental data indicates that the MamBo-3-Hydra-N3 hybrid model yields the lowest EERs on the ASV21LA (0.81%) and ITW (4.97%) datasets. Regarding aggregate performance, MamBo-3-Hydra-N3 achieves an average EER of 4.75%, representing a visible improvement over the 5.53% for the MamBo-4-Hydra-N1 configuration, ref in Table 2. In comparison to baselines, MamBo-3-Hydra-N3 demonstrates relative improvements of 12.90% and 9.57% over XLSR-Mamba on 21LA and 21DF, respectively. Furthermore, it registers a 15.04% improvement over Fake-Mamba (L) on the ITW dataset. Based on these empirical findings, and considering its parameter efficiency, MamBo-3-Hydra-N3 is selected as the primary backbone for the proposed XLSR-MamBo framework.

6 Conclusion

In this paper, we introduced XLSR-MamBo, a modular framework that integrates a pre-trained XLSR front-end with hybrid SSM-Attention back-ends. This design relies on complementary inductive biases: leveraging SSMs for efficient compression of local temporal artifacts and Attention mechanisms for the precise retrieval of global spectral inconsistencies. Through a systematic evaluation of topological designs and stacking depths, we explored the synergistic potential of this hybridization for ADD. Experimental results indicate that the MamBo-3-Hydra-N3 configuration achieves competitive performance on the ASV21LA, DF, and ITW datasets, comparable to or surpassing several state-of-the-art single systems with similar parameter counts. This performance validates the benefit of Hydra's native bidirectional modeling. Theoretically, the quasiseparable matrices employed by Hydra are strictly more expressive than the mixer matrices of heuristic addition-based bidirectional SSMs employed in prior works. This enables the model to capture more complex, holistic non-causal dependencies required to distinguish subtle spoofing artifacts without structural redundancy. Furthermore, evaluations on the DFADD dataset suggest that the framework possesses promising generalization capabilities against advanced diffusion- and flow-matching-based synthesis methods. Notably, our analysis on scaling properties reveals that increasing backbone depth is essential for mitigating the high performance variance and inference instability often observed in shallower models. Deeper configurations effectively enhance detection robustness across diverse generative algorithms. Collectively, these findings demonstrate that hybrid SSM-Attention architectures are promising alternatives to traditional Transformer-based or pure SSM-based backbones. Future work will focus on advancing these hybrid architectures to keep pace with the rapid evolution of speech generation technologies, with the aim of developing more effective and reliable tools for detecting emerging spoofing attacks.

Limitations

Despite the promising performance of XLSR-MamBo, this study acknowledges several limitations. First, the current evaluation relies exclusively

on the single-source training paradigm using the ASV19LA dataset. While this serves as a standard baseline, the absence of large-scale or mixed-source training scenarios limits our assessment of the framework's full scalability. Consequently, it remains unclear whether the reported performance is constrained by data diversity or architectural capacity, and the theoretical upper bound of the model's generalization capability across diverse distributions remains to be fully determined. Second, the datasets utilized in this work are predominantly English-centric. Although the XLSR front-end benefits from large-scale cross-lingual pre-training, the downstream fine-tuning was restricted to English speech. This potential linguistic bias leaves the model's robustness unverified against deepfakes in languages with distinct phonological structures or tonal characteristics, posing a risk for cross-lingual deployment. Third, we observed a rapid convergence phenomenon during training. Models frequently achieved minimal validation loss within very few epochs (< 10), triggering early stopping. We observed that the checkpoint with the lowest validation loss did not necessarily correspond to the best-performing model for unseen scenarios. Counterintuitively, these early-stage checkpoints demonstrated competitive generalization capabilities. The underlying optimization dynamics governing this relationship between rapid convergence and robust generalization remain an open question and warrant further investigation in future work.

Ethical Considerations

While this study utilizes public datasets, the real-world deployment of ADD systems could raise privacy concerns regarding data transmission. Beyond the obvious risks of uploading raw audio, transmitting intermediate representations also presents security vulnerabilities; feature inversion attacks can potentially reconstruct speaker identity and linguistic content from these embeddings. Consequently, future implementations should prioritize privacy-preserving architectures such as on-device processing that fundamentally exclude raw audio transmission and mitigate reconstruction risks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 6227134), Guangzhou Basic and Applied Basic Research Foundation (Grant No. 2025A04J3585). We use AI for writing assistance for paraphrasing and polishing the original content.

References

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. 2024. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1763–1840. PMLR.

Hussam Azzuni and Abdulmotaleb El Saddik. 2025. Voice cloning: Comprehensive survey. *arXiv preprint arXiv:2505.00579*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sarah Barrington, Emily A Cooper, and Hany Farid. 2025. People are poorly equipped to detect ai-powered voice clones. *Scientific Reports*, 15(1):11004.

Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, and 1 others. 2025. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982. ISCA.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sirui Chen, Jingji Chen, Siqi Zhu, Ziheng Jiang, Yanghua Peng, and Xuehai Qian. 2025. Mesh-attention: A new communication-efficient distributed attention with improved data locality. *arXiv preprint arXiv:2512.20968*.

Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Zhao Lv, and Cunhang Fan. 2024. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. In *Interspeech 2024*, pages 2720–2724. ISCA.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Phuong Tuan Dat and Tran Huy Dat. 2025. Xlsr-kanformer: A kan-intergrated model for synthetic speech detection. In *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, pages 1–6. IEEE.

Jiawei Du, I-Ming Lin, I-Hsiang Chiu, Xuanjun Chen, Haibin Wu, Wenze Ren, Yu Tsao, Hung-yi Lee, and Jyh-Shing Roger Jang. 2024. Dfadd: The diffusion and flow-matching based audio deepfake dataset. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 921–928. IEEE.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. 2024. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. arxiv 2023. In *The Conference On Language Modeling (COLM)*.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*. ISCA.

Sukjun Hwang, Aakash Lahoti, Ratish Puduppully, Tri Dao, and Albert Gu. 2024. Hydra: Bidirectional state space models through generalized matrix mixers. *Advances in Neural Information Processing Systems*, 37:110876–110908.

Gautam Siddharth Kashyap, Zohaib Hasan Siddiqui, Mohammad Anas Azeez, Rafiq Ali, Shantanu Kumar, Navin Kamuni, and Jiechao Gao. 2025. Fooling the forgers: A multi-stage framework for audio deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Awais Khan, Khalid Mahmood Malik, James Ryan, and Mikul Saravanan. 2023. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review*, 56(Suppl 1):513–566.

Yassine El Kheir, Tim Polzehl, and Sebastian Möller. 2025. Bicrossmamba-st: Speech deepfake detection with bidirectional mamba spectro-temporal cross-attention. *arXiv preprint arXiv:2505.13930*.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, and 1 others. 2025. Jamba: A hybrid transformer-mamba language model. In *The International Conference on Learning Representations (ICLR)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Xiaohui Liu, Meng Liu, Longbiao Wang, Kong Aik Lee, Hanyi Zhang, and Jianwu Dang. 2023. Leveraging positional-related local-global dependency for synthetic speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *The International Conference on Learning Representations (ICLR)*.

Joel R. McConvey. 2025. Deepfakes a 'now problem' as eu ai act passes compliance deadline: Reality defender. Last accessed: 2025-12-27.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? In *Interspeech 2022*. ISCA.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann,

Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024. Can mamba learn how to learn? a comparative study on in-context learning tasks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2025. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *The International Conference on Learning Representations (ICLR)*.

Eros Rosello, Alejandro Gómez Alanís, Angel M Gomez, Antonio M Peinado, N Harte, J Carson-Berndsen, and G Jones. 2023. A conformer-based classifier for variable-length utterance processing in anti-spoofing. In *Interspeech*, volume 2023, pages 5281–5285.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. 2023. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 904–912.

Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. 2022. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE.

Duc-Tuan Truong, Ruijie Tao, Tuan Nguyen, Hieu-Thi Luong, Kong Aik Lee, and Eng Siong Chng. 2024. Temporal-channel modeling in multi-head self-attention for synthetic speech detection. In *Interspeech 2024*, pages 537–541. ISCA.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, and 1 others. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Yang Xiao and Rohan Kumar Das. 2025. Xlsr-mamba: A dual-column bidirectional state space model for spoofing attack detection. *IEEE Signal Processing Letters*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International conference on machine learning*, pages 10524–10533. PMLR.

Xi Xuan, Zimo Zhu, Wenxin Zhang, Yi-Cheng Lin, and Tomi Kinnunen. 2025. Fake-mamba: Real-time speech deepfake detection using bidirectional mamba as self-attention's alternative. In *Proceedings of the IEEE ASRU*.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and 1 others. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2024. Gated delta networks: Improving mamba2 with delta rule. In *The International Conference on Learning Representations (ICLR)*.

Tianle Yang, Chengzhe Sun, Siwei Lyu, and Phil Rose. 2026. Forensic deepfake audio detection using segmental speech features. *Forensic Science International*, 379:112768.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qishan Zhang, Shuangbing Wen, and Tao Hu. 2024. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6765–6773.

Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2025. Mamba in speech: Towards an alternative to self-attention. *IEEE Transactions on Audio, Speech and Language Processing*.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

A Appendix

A.1 Datasets

- ASVspoof-2019-LA: `https://datashare.ed.ac.uk/handle/10283/3336`
- ASVspoof-2021-LA: `https://zenodo.org/record/4837263`
- ASVspoof-2021-DF: `https://zenodo.org/record/4835108`

- In-the-Wild: https://deepfake-total.com/in_the_wild
- DFADD: https://huggingface.co/datasets/isjwdu/DFADD

A.2 Pre-Trained Model

- XLSR-300M: https://docs.pytorch.org/audio/main/generated/torchaudio.pipelines.WAV2VEC2_XLSR_300M.html