Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

MASV: Speaker Verification with Global and Local Context Mamba

Yang Liu[1], *Li Wan*[1], *Yiteng Huang*[1], *Ming Sun*[1], *Xinhao Mei*[1], *Xubo Liu*[1], *Yangyang Shi*[1], *Florian Metze*[1]

[1]Meta, United States

yangliuav@gmail.com, wwanli@meta.com, yah@meta.com, sunming425@meta.com, xinhaomei@meta.com, xuboliu@meta.com, yyshi@meta.com, fmetze@meta.com

Abstract

Deep learning models like Convolutional Neural Networks and transformers have shown impressive capabilities in speech verification, gaining considerable attention in the research community. However, CNN-based approaches struggle with modeling long-sequence audio effectively, resulting in suboptimal verification performance. On the other hand, transformer-based methods are often hindered by high computational demands, limiting their practicality. This paper presents the MASV model, a novel architecture that integrates the Mamba module into the ECAPA-TDNN framework. By introducing the Local Context Bidirectional Mamba and Tri-Mamba block, the model effectively captures both global and local context within audio sequences. Experimental results demonstrate that the MASV model substantially enhances verification performance, surpassing existing models in both accuracy and efficiency.
Index Terms: speaker verification, Mamba module, sequential modeling

1. Introduction

Speech verification is essential for validating the authenticity and reliability of speech inputs, particularly in edge devices like virtual reality (VR) systems and smart glasses, where high security and privacy are critical. In such applications, the demand for both accurate and efficient verification is paramount. Recent advances in deep learning, especially using architectures like Convolutional Neural Networks (CNNs) and Transformers, have driven substantial progress, setting new standards for accuracy and efficiency in this field.

Despite recent advancements, current models still face significant challenges that limit their performance. Studies on CNN-based methods [1] demonstrate that while these models are effective at capturing local features, they struggle to model the full context of long-sequence audio signals because of their limited receptive fields. This limitation often results in suboptimal performance in complex speech verification tasks. In contrast, Transformer-based models [2] are adept at modeling long-range dependencies but are hindered by high computational complexity. The quadratic scaling of these models with sequence length reduces their practicality for real-time applications.

To overcome these challenges, recent research has explored large pre-trained models and advanced pooling mechanisms. For example, Attentive Statistics Pooling (ASP) and Progressive Channel Fusion (PCF) have proven effective in enhancing the robustness and accuracy of speaker embeddings [3, 4]. Additionally, the development of innovative loss functions, such as AAM-Softmax and its variants, has improved the discriminative power of embeddings, ensuring better separation of intra-class and inter-class features [5, 6].

State-space models (SSMs) [7, 8] have recently emerged as a promising approach to overcoming these limitations. SSMs establish long-range dependencies with linear computational complexity, providing an efficient solution for processing long sequences. Models like Mamba [9] and Mamba-2 [10] , which incorporate SSMs, have demonstrated their effectiveness across domains such as natural language processing [11, 12] and vision tasks [9, 13]. However, their potential in the field of speech verification has not been thoroughly explored.

In this paper, we propose the MASV model, a novel architecture that incorporates the Mamba-2 module within the ECAPA-TDNN framework for speaker verification. The primary innovation of this model is the introduction of the Local Context Bidirectional Mamba (LCB-Mamba) and the Tri-Mamba block, which together allow the model to capture both global and local contextual information in audio sequences. The LCB-Mamba improves performance in streaming applications by using local context without depending on future information, which offers a significant advantage over traditional bidirectional models. The Tri-Mamba block further enhances the model's capability by efficiently integrating both local and global context, surpassing the performance of conventional convolutional layers. This architecture overcomes the limitations of CNNs and Transformers, providing a more efficient and accurate solution for real-time speaker verification. Our experimental results show that MASV outperforms existing models, achieving notable improvements in both verification accuracy and computational efficiency.

2. Background: Mamba

State Space Models (SSMs) are designed to map a 1-dimensional function or sequence $x(t) \mapsto y(t) \in \mathbb{R}$ and can be represented through a linear Ordinary Differential Equation (ODE):

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t), \tag{1}$$
$$y(t) = \mathbf{C}x(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ are parameters and $x(t) \in \mathbb{R}^N$ represents the implicit latent state. The Structured State Space Sequence Models (S4) [7] enhance basic SSMs by structuring the state matrix $\mathbf{A}$ and implementing an efficient algorithm. The state matrix is initialised using the High-Order Polynomial Projection Operator (HIPPO) [14], allowing for deep sequence models with robust and efficient long-range reasoning capabilities. This approach has outperformed Transformers [15] on the Long Range Arena Benchmark [16].

10.21437/Interspeech.2025-82

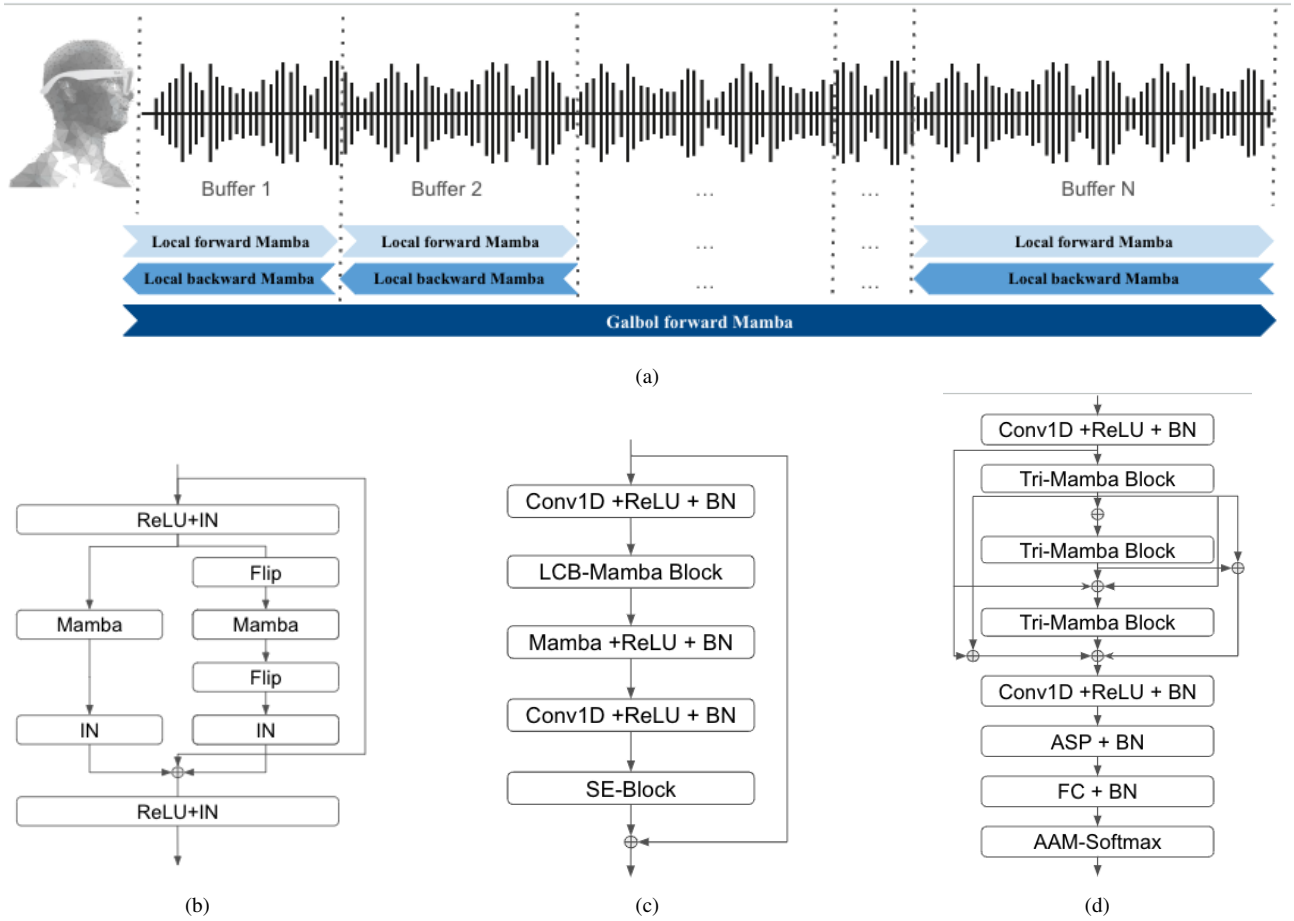


Figure 1: *MASV Pipeline and Architectures. (a) Streaming application pipeline where each audio buffer is processed by two independent Mamba layers for local context and a global Mamba layer for accumulated context. (b) Architecture of the LCB-Mamba Block. (c) Architecture of the Tri-Mamba Block, including the SE (Squeeze-and-Excitation) layer. (d) Full MASV architecture featuring the ASP (Attentive Statistics Pooling) layer.*

The Mamba and Mamba-2 [10] extends S4 for discrete data, such as text and genome sequences, with two primary innovations. First, it employs an input-dependent selection mechanism that dynamically adjusts the S4 parameters based on input data, differing from traditional time- and input-invariant S4. Second, Mamba introduces a hardware-aware algorithm that scales linearly with sequence length, enhancing computational efficiency on modern hardware. By integrating S4 blocks with linear layers, Mamba achieves superior performance across various long-sequence tasks, including text, speech, and genomics.

3. Proposed method

The proposed MASV model is built upon the ECAPA-TDNN framework, a state-of-the-art approach in speaker verification. Figure 1 presents the overall pipeline of the MASV model, highlighting components designed to address the limitations of traditional CNNs and Transformers in real-time applications. Specifically, we replace the Res2 Dilated CNN with the LCB-Mamba block to better handle local context within audio sequences. In addition, the Tri-Mamba block is introduced to effectively integrate both local and global context, and full-scale skip connections are implemented to facilitate the flow of information across different layers.

3.1. Local Context Bidirectional Mamba

In streaming speaker verification, audio data is received in segments known as audio buffers, which can have varying lengths. Traditional bidirectional models typically rely on future context for sequence modeling, which is not feasible in real-time applications where data arrives sequentially. To address this limitation, we propose the Local Context Bidirectional Mamba (LCB-Mamba) block, designed to incorporate local context information within each buffer without depending on future data.

The LCB-Mamba block enhances the model's performance by integrating local context directly into the state computation. For the local forward Mamba, the state $\overrightarrow{\mathbf{A}}x(t)$ at time step t depends not only on the current input $x(t)$ and the previous hidden state $\overrightarrow{\mathbf{A}}x(t-1)$, but also on a local context window of inputs:

$$\overrightarrow{\mathbf{A}}x(t) = \text{Mamaba}(\overrightarrow{\mathbf{A}}x(t-1), \overrightarrow{\mathbf{c}}) \tag{3}$$

where $\overrightarrow{\mathbf{c}}$ represents the forward local context in buffer. Similarly, the hidden state $\overleftarrow{\mathbf{A}}x(t)$ is computed by:

$$\overleftarrow{\mathbf{A}}x(t) = \text{Mamaba}(\overleftarrow{\mathbf{A}}_{t+1}, \overleftarrow{\mathbf{c}}) \tag{4}$$

The final state representation is obtained by concatenating the forward and backward states: $\mathbf{A}x(t) = [\overrightarrow{\mathbf{A}}x(t); \overleftarrow{\mathbf{A}}x(t)]$.

The LCB-Mamba architecture, as shown in Fig. 1b, begins with the input passing through a Rectified Linear Unit (ReLU) activation function combined with Instance Normalization (IN). The network then branches into two parallel paths: the forward path consists of a Mamba layer followed by an IN layer, while the backward path includes two Flip layers interspersed with Mamba layers and IN. The outputs from both paths, along with the initial input, are merged and processed through another ReLU layer with IN.

By effectively incorporating local context information, the LCB-Mamba architecture is optimized to capture local dependencies within the buffer, particularly in streaming applications where future context is not available. This design ensures robustness and efficiency in real-time processing scenarios that require high adaptability.

3.2. Tri-Mamba Block

While the LCB-Mamba block effectively captures local context, it lacks a mechanism to connect different audio buffers to form a more comprehensive global context. To address this limitation, we propose the Tri-Mamba block, which integrates both local and global context information more efficiently.

As illustrated in Fig. 1c, the Tri-Mamba block replaces the SE-Res2Block in the original ECAPA-TDNN architecture. The block begins with a Conv1D layer, followed by ReLU activation and Batch Normalization (BN), which extracts essential features from the input. The output then passes through the LCB-Mamba block, which captures bidirectional local context using the mechanisms. To further enhance feature extraction, an additional Mamba block is incorporated with ReLU activation and BN to model more complex patterns. This is followed by another Conv1D layer with ReLU and BN to refine the extracted features. Finally, a Squeeze-and-Excitation (SE) block recalibrates the feature representation by adjusting the channel-wise feature responses.

The combination of these components in the Tri-Mamba block allows for a more comprehensive modeling of audio sequences, effectively leveraging both local and global context information. This design enhances the overall performance of the MASV model, particularly in scenarios requiring real-time speaker verification.

3.3. MASV with full-scale skip

To improve information flow and feature retention, we introduce full-scale skip connections in the MASV model. These connections perform element-wise additions between input features and outputs of corresponding layers (Fig. 1d), preserving both low-level and high-level features. This design mitigates issues like gradient vanishing in deep networks, enhancing training efficiency and convergence.

By combining detailed local information with global context, full-scale skip connections improve accuracy and computational efficiency, making the MASV model robust for real-time speaker verification.

4. Experiments

4.1. Dataset

We use a geographically filtered private development set as our dataset, consisting of 5,182,021 utterances from 28,622 speakers. This dataset was recorded in a soundproof studio to ensure controlled noise conditions and high-quality audio samples (48 kHz). Our approach addresses the limitations of VoxCeleb2, including noise contamination, uncontrolled recording environments, and privacy concerns, by providing consistent data quality and geographically restricting participation. The input features are 80-dimensional log-Mel filter bank coefficients, normalised via cepstral mean normalisation without voice activity detection.

Table 1: *Performance on Evaluation Sets*

Model	EER	DCF
ResNet18	1.610	0.1989
ResNet34	1.264	0.1341
ECAPA (C=512)	1.158	0.1221
ECAPA (C=1024)	0.956	0.1266
PCF-ECAPA (C=512)	0.818	0.1058
PCF-ECAPA (C=1024)	0.801	0.1092
MASV (C=512)	**0.805**	**0.1025**
MASV (C=1024)	**0.795**	**0.0998**

Table 2: *Ablation Study of the MASV Architecture*

Configuration	Params	EER	minDCF
Base Model (512 channels)	6.2	1.158	0.1221
+ LCB-Mamba Block	8.7	0.895	0.0982
+ Tri-Mamba Block	9.1	0.840	0.0947
Complete MASV	9.2	**0.795**	**0.0875**

4.2. Model

Our baseline model is implemented with 512 and 1024 channels, while the proposed MASV (Mamba-based Speaker Verification) model adopts the same configuration with additional enhancements. The MFA module output channels are fixed at 1536 to maintain consistency. For comparison, ResNet18, ResNet34 architectures and PCF-ECAPA [2], are also evaluated using the same attentive statistics pooling mechanism. Training is performed using the Adam optimizer with a cyclical learning rate strategy over 100k-step cycles, varying from 1e-8 to 1e-3, and a weight decay of 5e-5. We employ a batch size of 256 and the Circle loss function with $m = 0.35$ and $s = 60$ to enforce stronger constraints on speaker embeddings. All models are trained under these same conditions.

4.3. Evaluation

We report performance metrics in terms of Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $p_{\text{target}} = 0.01$, $C_{\text{FA}} = C_{\text{Miss}} = 1$.

Table 1 shows that the MASV model, configured with 512 and 1024 channels, outperforms the baseline ECAPA-TDNN and ResNet models, achieving an average relative improvement of 15.6% in EER and 15.2% in minDCF. This confirms the effectiveness of our enhancements in improving speaker verification performance.

The ablation study in Table 2 demonstrates the impact of each component in the MASV architecture. The addition of the LCB-Mamba block reduces EER by 22.7%, showing its strength in capturing local context without future information. The Tri-Mamba block further enhances performance by integrating both local and global context, while full-scale skip connections improve information flow and stability. The complete MASV model achieves the lowest EER and minDCF, validat-

Table 3: *Performance on VoxCeleb Official Evaluation Sets*

Stage		Vox1O		Vox1E		Vox1H		Vox20-dev		Vox21-dev		Vox22-dev	
		EER	$DCF_{0.01}$	EER	$DCF_{0.01}$	EER	$DCF_{0.01}$	EER	$DCF_{0.01}$	EER	$DCF_{0.01}$	EER	$DCF_{0.01}$
ResNet18		1.510	0.1789	1.559	0.1760	2.679	0.2642	4.235	0.3540	4.835	0.3798	3.624	0.4142
ResNet34		1.164	0.1141	1.167	0.1285	2.099	0.2127	3.365	0.2794	3.806	0.3052	2.825	0.3062
ECAPA	(C=512)	1.058	0.1021	1.205	0.1371	2.182	0.2155	3.557	0.2990	4.045	0.3247	2.930	0.3541
	(C=1024)	0.856	0.1066	1.074	0.1285	2.009	0.2021	3.365	0.2725	4.142	0.3353	2.830	0.3110
PCF-	(C=512)	0.718	0.0858	0.792	0.1138	1.802	0.1750	2.959	0.2250	3.684	0.3073	2.630	0.2836
ECAPA	(C=1024)	0.718	0.0892	0.891	0.1024	1.707	0.1754	2.831	0.2339	3.527	0.2880	2.333	0.2666
MASV-	(C=512)	0.655	0.0775	0.725	0.1062	1.702	0.1633	2.634	0.2015	3.530	0.2850	2.450	0.2500
Mamba1	(C=1024)	0.630	0.0755	0.731	0.0931	1.622	0.1587	2.430	0.1955	3.300	0.2700	2.200	0.2450
MASV-	(C=512)	**0.510**	**0.0593**	**0.570**	**0.0836**	**1.350**	**0.1278**	**2.080**	**0.1592**	**2.800**	**0.2240**	**1.936**	**0.1960**
Mamba2	(C=1024)	**0.436**	**0.0560**	**0.507**	**0.0721**	**1.266**	**0.1226**	**1.862**	**0.1516**	**2.600**	**0.2080**	**1.720**	**0.1880**

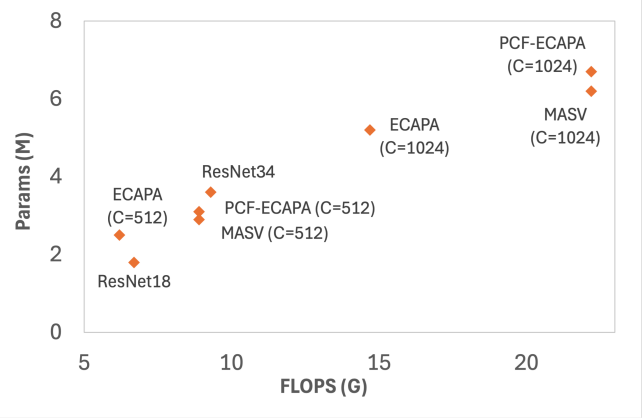


Figure 2: *Computational Complexity Comparison on*

ing the combined contribution of these components to a more robust speaker verification solution.

4.4. Computational Complexity Experiment

To evaluate the efficiency of the MASV model compared to the baseline models, we analyze their computational complexity by estimating the number of floating-point operations per second (FLOPS) required for each model configuration. Fig. 2 summarizes the number of parameters and FLOPS for each model, including ResNet, ECAPA, PCF-ECAPA, and our proposed MASV.

The MASV model, with both 512 and 1024 channel configurations, demonstrates a balance between parameter size and computational cost. For instance, MASV (C=1024) offers lower FLOPS (6.2G) than PCF-ECAPA (C=1024) while maintaining a competitive parameter count (22.2M), indicating its efficiency in real-time applications.

4.5. Experimental Validation on Publicly Available Datasets

To validate the effectiveness of the proposed MASV-Mamba model, we conducted experiments on publicly available VoxCeleb evaluation sets, including VoxCeleb-O, VoxCeleb-E, VoxCeleb-H, and multiple VoxCeleb2 development sets. For fair comparison, all training dataset is public dataset. Although our model is build with Mamba-2, we also eval the performance with Mamba for comparison. Table 3 presents the detailed results, where MASV-Mamba2 ($C = 1024$) consistently achieves the lowest Equal Error Rates (EER) and Detection Cost Function ($DCF_{0.01}$) values across all datasets, demonstrating its superior performance in speaker verification tasks.

Specifically, on the VoxCeleb1-O evaluation set, MASV-Mamba2 ($C = 1024$) achieves an EER of 0.436% and a $DCF_{0.01}$ of 0.0560, significantly outperforming baseline models such as ResNet18, ResNet34, and ECAPA-TDNN. Similar trends are observed on VoxCeleb1-E and VoxCeleb1-H, where the proposed model achieves the best performance with EERs of 0.507% and 1.266%, respectively. The corresponding $DCF_{0.01}$ values of 0.0721 and 0.1226 further highlight the model's robustness under strict decision-making thresholds.

On the VoxCeleb2 development sets, MASV-Mamba2 ($C = 1024$) maintains its superior performance, achieving EERs of 1.862%, 2.600%, and 1.720% on Vox20-dev, Vox21-dev, and Vox22-dev, respectively. Additionally, it achieves the lowest $DCF_{0.01}$ values across these datasets, demonstrating its consistent effectiveness even in complex and challenging evaluation scenarios.

5. Conclusion

This paper introduces the MASV model, a novel speaker verification architecture that enhances real-time accuracy and efficiency by integrating the LCB-Mamba block, Tri-Mamba block, and full-scale skip connections. These innovations address the limitations of existing models by improving local and global context capture and ensuring robust information flow. Our results show that MASV significantly reduces EER and minDCF compared to baseline models while maintaining computational efficiency, making it an effective solution for dynamic, real-world environments.

6. References

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[2] Z. Zhao, Z. Li, W. Wang, and P. Zhang, "Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] T. Liu, K. A. Lee, Q. Wang, and H. Li, "Golden gemini is all you need: Finding the sweet spots for speaker verification,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[4] H. Wang, H. Li, and B. Li, "Vot: Revolutionizing speaker verification with memory and attention mechanisms," *arXiv preprint arXiv:2312.16826*, 2023.

[5] K. A. Al-Karawi, "Real-time adaptive training for forensic speaker verification in reverberation conditions," *International Journal of Speech Technology*, vol. 26, no. 4, pp. 1079–1089, 2023.

[6] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 513–566, 2023.

[7] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[8] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[10] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," *arXiv preprint arXiv:2405.21060*, 2024.

[11] M. Pióro, K. Ciebiera, K. Król, J. Ludziejewski, and S. Jaszczur, "Moe-mamba: Efficient selective state space models with mixture of experts," *arXiv preprint arXiv:2401.04081*, 2024.

[12] Z. Yang, A. Mitra, S. Kwon, and H. Yu, "Clinicalmamba: A generative clinical language model on longitudinal clinical notes," *arXiv preprint arXiv:2403.05795*, 2024.

[13] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[14] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," *arXiv preprint arXiv:2011.04006*, 2020.