A Survey of Mamba

HAOHAO QU, The Hong Kong Polytechnic Univeristy, China

LIANGBO NING, The Hong Kong Polytechnic Univeristy, China

RUI AN, The Hong Kong Polytechnic Univeristy, China

WENQI FAN*, The Hong Kong Polytechnic Univeristy, China

TYLER DERR, Vanderbilt University, USA

XIN XU, The Hong Kong Polytechnic Univeristy, China

QING LI, The Hong Kong Polytechnic Univeristy, China

Deep learning, as a vital technique, has sparked a notable revolution in artificial intelligence (AI), resulting in a great change in human lifestyles. As the most representative architecture, Transformers have empowered numerous advanced models, especially the large language models (LLMs) that comprise billions of parameters, becoming a cornerstone in deep learning. Despite the impressive achievements, Transformers still face inherent limitations, particularly the time-consuming inference resulting from the quadratic computation complexity of attention calculation. Recently, a novel architecture named **Mamba**, drawing inspiration from classical state space models, has emerged as a promising alternative for building foundation models, delivering comparable modeling abilities to Transformers while preserving near-linear scalability concerning sequence length. This has sparked an increasing number of studies actively exploring Mamba's potential to achieve impressive performance across diverse domains. Given such rapid evolution, there is a critical need for a systematic review that consolidates existing Mamba-empowered models, offering a comprehensive understanding of this emerging model architecture. In this survey, we therefore conduct an in-depth investigation of recent Mamba-associated studies, covering from three main aspects: *the advancements of Mamba-based models*, *the techniques of adapting Mamba to diverse data*, and *the applications where Mamba can excel*. Specifically, we first recall the foundational knowledge of various representative deep learning models and the details of Mamba-1&2 as preliminaries. Then, to showcase the significance of Mamba for AI, we comprehensively review the related studies focusing on Mamba models' architecture design, data adaptability, and applications. Finally, we present an discussion of current limitations and explore various promising research directions to provide deeper insights for future investigations.

Additional Key Words and Phrases: State Space Model, Mamba, Sequence Modeling, Foundation Models, Language Models

---

*Corresponding author, Wenqi Fan, Department of Computing (COMP) & Department of Management and Marketing (MM).

Authors' Contact Information: Haohao Qu, haohao.qu@connect.polyu.hk, The Hong Kong Polytechnic Univeristy, Hong Kong, China; Liangbo Ning, liangbo1123.ning@connect.polyu.hk, The Hong Kong Polytechnic Univeristy, Hong Kong, China; Rui An, rui77.an@connect.polyu.hk, The Hong Kong Polytechnic Univeristy, Hong Kong, China; Wenqi Fan, wenqifan03@gmail.com, The Hong Kong Polytechnic Univeristy, Hong Kong, China; Tyler Derr, Tyler.Derr@vanderbilt.edu, Vanderbilt University, Nashville, USA; Xin Xu, xin.xu@polyu.edu.hk, The Hong Kong Polytechnic Univeristy, Hong Kong, China; Qing Li, qing-prof.li@polyu.edu.hk, The Hong Kong Polytechnic Univeristy, Hong Kong, China.

---

1 Introduction

Over the past two decades, deep learning, as the most prominent artificial intelligence (AI) technique, has brought about a revolution in various domains such as healthcare [88], autonomous systems [37, 60], recommender systems [104, 228], and financial services [144, 216]. This period has witnessed the emergence of numerous deep neural networks (DNNs) that have significantly altered human lifestyles, offering immense convenience to individuals. One notable example is U-Net [151, 164], a robust deep learning model within the field of vision, which is extensively employed in medical imaging for the examination of radiology scans like MRI and CT scans. Its application assists in the identification and diagnosis of diseases, showcasing its effectiveness in this critical healthcare domain [112, 191]. Moreover, Graph Neural Networks (GNNs) are employed in handling graph-structure data to support smart services, such as recommender systems that suggest personalized content, products, or services to users [41, 42, 194]. Furthermore, Recurrent Neural Networks (RNNs) are extensively adopted in machine translation due to their ability to capture the sequential and contextual information essential for accurate translations [119, 167], empowering individuals from diverse linguistic backgrounds to effectively communicate and comprehend each other's ideas, opinions, and information.

Among the various deep learning architecture, Transformers have recently stood out and established its dominance across a broad spectrum of applications [34, 175]. For instance, as the most representative large foundation models, large language models (LLMs) like ChatGPT are fundamentally built on the Transformer architecture [2, 146, 228]. By scaling their model sizes to billions and training on a mix of diverse data sources, these Transformer-based models have demonstrated human-level intelligence with its impressive capabilities in language understanding, common sense reasoning, and in-content-learning [44, 217]. This remarkable success is bolstered by the attention mechanism [174], which enables the Transformer-based models to concentrate on relevant parts of the input sequence and facilitate better contextual understanding. However, the attention mechanism also introduces a significant computational overhead that increases quadratically with the input size [124, 233], which presents challenges in processing lengthy inputs. For example, the rapid growth in computational cost makes Transformers impractical or infeasible to process substantial sequences, thereby limiting their applicability in tasks like document-level machine translation [131] or long document summarization [94].

Recently, a promising architecture, structured state space sequence models (SSMs) [58], have emerged to efficiently capture complex dependencies in sequential data, becoming a formidable competitor to Transformer. These models, inspired by classical state space models [89], can be considered as a fusion of recurrent neural networks and convolutional neural networks. They can be computed efficiently using either recurrence or convolution operations, achieving linear or near-linear scaling with sequence length, thus significantly reducing the computational costs. More specifically, as one of the most successful SSM variants, **Mamba** achieves comparable modeling capabilities to Transformers while maintaining linear scalability with sequence length [55], propelling it into the realm of focal topics. Mamba first introduces a simple yet effective selection mechanism that enables the model to filter out irrelevant information while retaining necessary and relevant data indefinitely by parameterizing the SSM parameters based on the input. Then, Mamba proposes a hardware-aware algorithm to compute the model recurrently with a scan instead of convolution, achieving up to 3×faster computation on A100 GPUs. As shown in Figure 1, the powerful modeling capabilities for complex and lengthy sequential data, along with near-linear scalability, position Mamba as an emerging foundation model, poised to revolutionize various domains of research and applications, such as computer vision [199, 234], natural language processing [111, 226], healthcare [152, 179, 198], etc. For example, Zhu et al. [234] propose Vim, which is 2.8×faster than DeiT [173] and saves 86.8% GPU memory when extracting features for high-resolution images. Dao and

Gu [29] show the connections between SSMs and variants of attention and propose a new architecture that refines selective SSM, achieving 2-8× faster on language modeling.
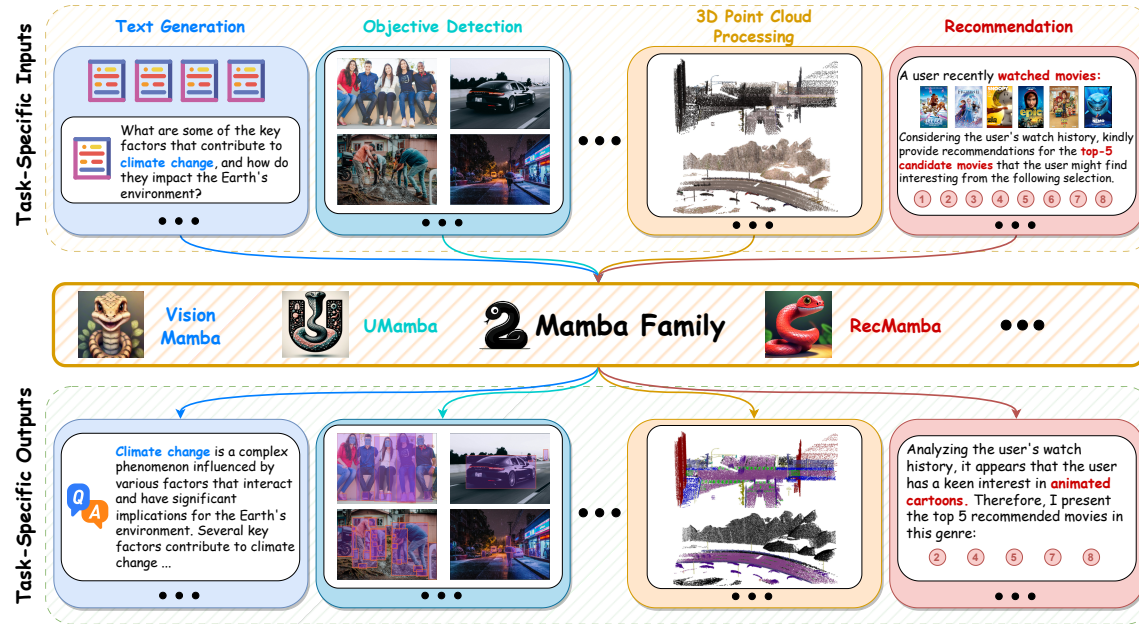

Fig. 1. Examples of the applications of Mamba-based models for different downstream tasks.

Motivated by the powerful long-sequence modeling capabilities of Mamba and its great efficiency, a substantial body of literature has emerged, focusing on employing and improving Mamba on various downstream tasks. Given this significant surge in studies related to Mamba, it is crucial to conduct a comprehensive review of existing literature and deliberate on potential directions for future research. In this survey, we thus conduct a comprehensive review of Mamba from several perspectives to provide newcomers with a fundamental understanding of Mamba's inner workings while helping experienced practitioners stay abreast of its latest developments. Specifically, the remaining survey is organized as follows: In Section 2, we recall the background knowledge of various representative deep neural networks, including RNNs, Transformers, and State Space Models, while the details of Mamba are introduced in Section 3. Subsequently, we summarize the recent advancements in Mamba-based studies from the perspectives of block design, scanning mode, and memory management in Section 4. Then, Section 5 presents the techniques of adapting Mamba to diverse data, including sequential and non-sequential data. Besides, representative applications of Mamba models are introduced in Section 6, while the challenges and future directions are presented in Section 7. Finally, we summarize the whole survey in Section 8.

Concurrent with our survey, several related surveys have been released, purely focusing on state space models [137, 183] and Vision Mamba [120, 199, 218]. Diverging from these surveys, this paper is centered on the associated research concerning Mamba. It systematically analyzes existing literature from a novel standpoint to explore *the evolution of Mamba architecture* and *the data adaptation methods utilized in Mamba-based models*.

2 Preliminary

Mamba is deeply intertwined with the recurrent framework of Recurrent Neural Networks (RNNs), the parallel computation and attention mechanism of Transformers, and the linear property of State Space Models (SSMs). Therefore, this section aims to present an overview of these three prominent architectures, serving as a prerequisite for comprehending Mamba thoroughly.

2.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) excel in processing sequential data due to their capability to retain internal memory [54]. Such networks have demonstrated remarkable effectiveness in a multitude of tasks that involve analyzing and predicting sequences, e.g., speech recognition, machine translation, natural language processing, and time-series analysis [69, 169]. In order to grasp the foundations of recurrent models, this section will offer a brief overview of the standard RNN formulation.

Specifically, at each discrete time step k, the standard RNN specifically processes a vector $x_k \in \mathbb{R}^D$ along with the previous step's hidden state $h_{k-1} \in \mathbb{R}^N$ to produce an output vector $o_k \in \mathbb{R}^O$ and update the hidden state to $h_k \in \mathbb{R}^N$. The hidden state serves as the network's memory and retains information about the past inputs it has seen. This dynamic memory allows RNNs to process sequences of varying lengths. Formally, it can be written as

$$h_k = \tanh(\mathbf{W}_{hx} x_k + \mathbf{W}_{hh} h_{k-1} + b_h), \tag{1}$$

$$o_k = \mathbf{W}_{oh} h_k + b_o, \tag{2}$$

where $\mathbf{W}_{hx} \in \mathbb{R}^{N \times D}$ is the weight matrix responsible for processing model inputs into hidden states, $\mathbf{W}_{hh} \in \mathbb{R}^{N \times N}$ is the recurrent connections between hidden states, $\mathbf{W}_{oh} \in \mathbb{R}^{O \times N}$ represents the weight used to generate outputs derived from hidden states, $b_h \in \mathbb{R}^N$ and $b_o \in \mathbb{R}^O$ correspond the biases, and tanh denotes the hyperbolic tangent activation function introducing non-linearity to the RNN model. In other words, RNNs are nonlinear recurrent models that effectively capture temporal patterns by harnessing the historical knowledge stored in hidden states.

However, there are several limitations associated with RNNs. First, RNNs have a restricted capability to effectively extract long-range dynamics within input sequences. As information traverses through successive time steps, the repeated multiplication of weights in the network can lead to dilution or loss of information. Consequently, it becomes challenging for RNNs to retain and recall information from earlier time steps while making predictions. Second, RNNs process sequential data incrementally, restricting their computational efficiency since each time step relies on the preceding one. This makes parallel computations challenging for them. Furthermore, conventional RNNs lack built-in attention mechanisms, which allow the network to capture global information within input sequences. This absence of attention mechanisms hinders the network's ability to selectively model the crucial segments of the data. o overcome these constraints, Transformers and State Space Models have emerged, each tackling these challenges from different perspectives. These two approaches will be further elaborated upon in the subsequent subsections.

2.2 Transformers

The Transformer [174] is a groundbreaking model in the realm of deep learning, revolutionizing various artificial intelligence applications. Its introduction marked a significant departure from traditional sequence-to-sequence models by employing a **self-attention** mechanism, facilitating the capture of global dependencies within model inputs. For instance, in natural language processing, this self-attention capability allows the model to comprehend relationships

between various positions in a sequence. It achieves this by assigning weights to each position based on its significance relative to other positions. More specifically, a sequence of input vectors $\mathbf{x}$ is first transformed into three types of vectors: **Query** Q, **Key** K, and **Value** V by utilizing linear transformations of the original input, defined by:

$$Q = \mathbf{x} \cdot \mathbf{W}^Q, K = \mathbf{x} \cdot \mathbf{W}^K, V = \mathbf{x} \cdot \mathbf{W}^V, \tag{3}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are the trainable parameters. The attention scores are computed by calculating the dot product of Q and K, then scaling the result by $\sqrt{d_K}$, where d_K is the dimension of the key vectors. Such procedures are then passed through a Softmax function to normalize the scores S and produce attention weights, defined by:

$$S = \text{Softmax}(\frac{QK^T}{\sqrt{d_K}})V, \tag{4}$$

Apart from performing a single attention function, multi-head attention is introduced to enhance the model's ability to capture different types of relationships and provide multiple perspectives on the input sequence. In multi-head attention, an input sequence is processed in parallel by multiple sets of self-attention modules. Each head operates independently, performing the exact computations as in the standard self-attention mechanism. The attention weights from each head are then combined to create a weighted sum of the value vectors. This aggregation step allows the model to leverage information from multiple heads and capture diverse patterns and relationships present in the input sequence. Mathematically, the multi-head attention is computed as follows:

$$\text{MultiHead}(Q, K, V) = (S_1 \oplus S_2 \dots \oplus S_m) \cdot \mathbf{W}^O,$$
$$\text{where } S_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_K}})V_i, \ i \in [1, m], \tag{5}$$

where m is the number of attention heads. $\oplus$ is the concatenation operation, and $\mathbf{W}^O$ is the linear transformation to project the multi-head attention scores to the final values.

2.3 State Space Models

State Space Models (SSMs) are a traditional mathematical framework utilized to depict the dynamic behavior of a system over time [89]. Recent years have found the widespread applications of SSMs in diverse fields like control theory, robotics, and economics [58, 59]. At its core, SSMs embody the system's behavior through a collection of hidden variables referred to as "states", enabling it to capture temporal data dependencies effectively. Different from RNNs, SSMs are linear models characterised by their associative property. To be specific, in a classical state space model, two key equations are formulated, i.e., state equation and observation equation, to model the relationships between input $x(t) \in \mathbb{R}$ and output $y(t) \in \mathbb{R}$ at current time t, through a N-dimensional hidden state $h(t) \in \mathbb{R}^N$. The process can be written by

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \tag{6}$$

$$y(t) = \mathbf{C}h(t) + \mathbf{D}x(t), \tag{7}$$

where $h'(t)$ is the derivative of current state $h(t)$, $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the state transition matrix that describes how states change over time, $\mathbf{B} \in \mathbb{R}^{N \times 1}$ is the input matrix that controls how inputs affect state changes, $\mathbf{C} \in \mathbb{R}^{1 \times N}$ denotes the output matrix that indicates how outputs are generated based on current states, and $\mathbf{D} \in \mathbb{R}$ represents the command

coefficient that determines how inputs affect outputs directly. In general, most SSMs exclude the second term in the observation equation, i.e., set $\mathbf{D}x(t) = 0$, which can be recognized as a skip connection in deep learning models.

2.3.1 **Discretization**. To adhere to the requirements of machine learning settings, SSMs must undergo a process of discretization that transforms continuous parameters into discrete parameters. Discretization methods generally aim to partition continuous time into K discrete intervals with equal integration area as possible. To achieve the goal, as one of the most representative solutions, Zero-Order Hold (ZOH) [138, 223] is successfully employed in SSMs, which assumes that the function value remains constant over the interval $\Delta = [t_{k-1}, t_k]$. After ZOH discretization, the SSM equations can be rewritten by

$$h_k = \overline{\mathbf{A}}h_{k-1} + \overline{\mathbf{B}}x_k, \tag{8}$$

$$y_k = \mathbf{C}h_k, \tag{9}$$

where $\overline{\mathbf{A}} = \exp(\Delta\mathbf{A})$, and $\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}$, k is the discrete time step. From these formulas, it is clear that the discrete SSM has a similar structure to recurrent neural networks and, therefore, discrete SSMs can accomplish inference processes with higher efficiency compared to Transformer-based models that compute attention on all inputs in each auto-regressive decoding iteration.

2.3.2 **Convolutional Computation**. The discrete SSM, being a linear system, possesses the associated property and therefore integrates seamlessly with convolutional computation. More specifically, it can calculate the output at each time step independently as follows:

$$y_0 = \mathbf{C}\overline{\mathbf{A}}^0\overline{\mathbf{B}}x_0, \tag{10}$$

$$y_1 = \mathbf{C}\overline{\mathbf{A}}^1\overline{\mathbf{B}}x_0 + \mathbf{C}\overline{\mathbf{A}}^0\overline{\mathbf{B}}x_1, \tag{11}$$

$$y_2 = \mathbf{C}\overline{\mathbf{A}}^2\overline{\mathbf{B}}x_0 + \mathbf{C}\overline{\mathbf{A}}^1\overline{\mathbf{B}}x_1 + \mathbf{C}\overline{\mathbf{A}}^0\overline{\mathbf{B}}x_2, \tag{12}$$

$$...... \tag{13}$$

$$y_k = \mathbf{C}\overline{\mathbf{A}}^k\overline{\mathbf{B}}x_0 + \mathbf{C}\overline{\mathbf{A}}^{k-1}\overline{\mathbf{B}}x_1 + ... + \mathbf{C}\overline{\mathbf{A}}^1\overline{\mathbf{B}}x_{k-1} + \mathbf{C}\overline{\mathbf{A}}^0\overline{\mathbf{B}}x_k. \tag{14}$$

By creating a set of convolutional kernels $\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, ..., \mathbf{C}\overline{\mathbf{A}}^k\overline{\mathbf{B}}, ...)$, the recurrent computation can be converted to a convolutional form as:

$$\mathbf{y} = \mathbf{x} * \overline{\mathbf{K}}, \tag{15}$$

where $\mathbf{x} = [x_0, x_1, ...]$ and $\mathbf{y} = [y_0, y_1, ...] \in \mathbb{R}^L$ denote the input and output sequences, respectively, while L is the sequence length. This convolutional computation allows SSMs to take full advantage of modern matrix computation hardware (e.g., GPUs) to enable parallel computing during training process, which is impossible with RNNs utilizing nonlinear activation functions. Notably, given an input $x(k)$ with D dimensions, the SSM computation will be calculated separately for each dimension to produce a D-dimensional output $y(t)$. In this case, the input matrix $\mathbf{B} \in \mathbb{R}^{N \times D}$, the output matrix $\mathbf{C} \in \mathbb{R}^{D \times N}$, and the command matrix $\mathbf{D} \in \mathbb{R}^{D \times D}$, while the state transition matrix remains unchanged, i.e., $\mathbf{A} \in \mathbb{R}^{N \times N}$.

2.3.3 **Relationship among RNN, Transformer, and SSM**. The computation algorithms of RNN, Transformer, and SSM are depicted in Figure 2. On the one hand, the conventional RNN operates within a non-linear recurrent

framework where each computation depends solely on the previous hidden state and the current input. While this format allows RNNs to quickly generate outputs during auto-regressive inference, it hampers their ability to fully exploit GPU parallel computing, leading to slower model training. On the other hand, the Transformer architecture performs matrix multiplications in parallel across multiple query-key pairs, which can be efficiently distributed across hardware resources, which enables faster training of attention-based models. However, when it comes to generating responses or predictions from Transformer-based models, the inference process can be time-consuming. For instance, the auto-regressive design of language models entails generating each token in the output sequence sequentially, which requires repetitive calculations of attention scores at each step, leading to slower inference times. Unlike RNNs and Transformers, which are limited to supporting only one type of computation, discrete SSMs have the flexibility to support both recurrent and convolutional computations, given its linear property. This unique capability allows SSMs to achieve not only efficient inference but also parallel training. However, it should be noted that the most conventional SSMs are time-invariant, meaning that their $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and Δ are unrelated to the model input x. This would limit context-aware modeling, which leads to inferior performance of SSMs in certain tasks such as selective copying [55].
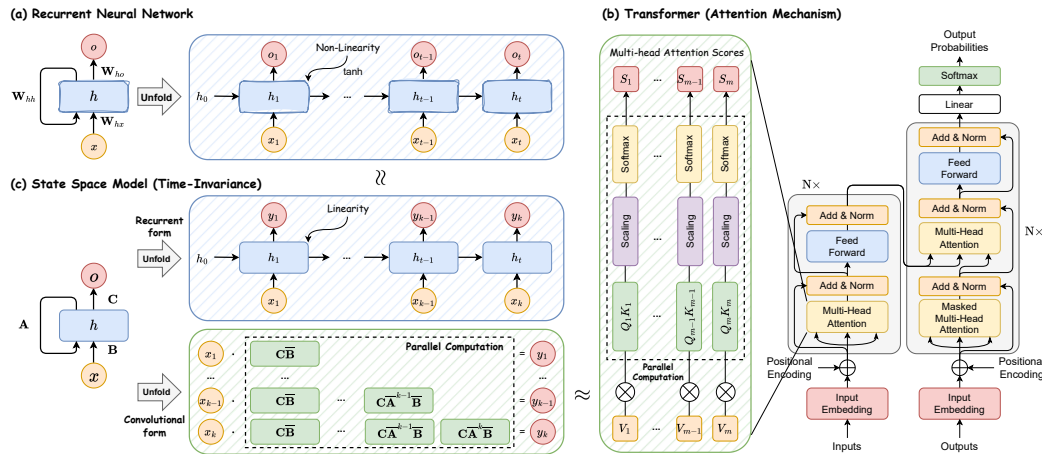

Fig. 2. An illustration of representative model architectures, namely Recurrent Neural Network (RNN), Transformer, and State Space Model (SSM). (a) RNNs function within a nonlinear recurrent framework, facilitating rapid outputs during auto-regressive inference. (b) Transformers execute matrix multiplications concurrently across numerous query-key pairs, facilitating parallel training. (c) SSMs exhibit versatility by accommodating both recurrent and convolutional computations due to their linear nature. This fusion harnesses the strengths of RNNs and Transformers, allowing SSMs for recurrent inference and parallel training. Despite this, traditional time-invariant SSMs fall short in context-aware modeling, resulting in diminished performance in specific tasks.

3 Mamba

To address the aforementioned drawback of traditional SSMs in terms of their inferior context-aware capabilities, **Mamba** is proposed by [55] as a potential alternative that promises to be a general sequence foundation model backbone. More recently, **Mamba-2** [29] proposes Structured Space-State Duality (SSD) that establishes a robust theoretical framework connecting structured SSMs and various forms of attention, allowing us to transfer algorithmic and systems optimizations originally developed for Transformers to SSMs. In this section, we will give a concise and clear introduction to Mamba and Mamba-2.

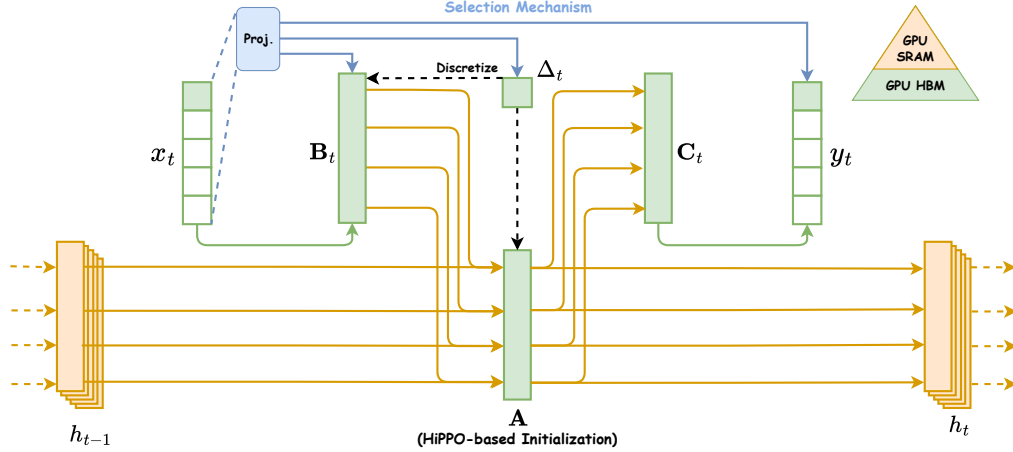### 3.1 Mamba-1: Selective State Space Model with Hardware-aware Algorithms

Fig. 3. Overview of the Selective State Space Model with hardware-aware state expansions. The selective mechanism introduces input-dependent dynamics, while the hardware computation algorithm optimizes memory usage within the more efficient levels of GPU memory hierarchy.

Conventional SSMs have shown limited effectiveness in modeling text and other information-dense data [55], impeding their progress in deep learning. In the pursuit of empowering SSMs with Transformers' modeling capabilities, Gu and Dao [55] introduces three innovative techniques based on Structured State Space Models, i.e., High-order Polynomial Projection Operator (HiPPO)-based Memory Initialization, Selection Mechanism, and Hardware-aware Computation, as illustrated in Figure 3. These techniques aim to enhance the capabilities of SSMs in long-range linear-time sequence modeling. In particular, the initialization strategy establishes a coherent hidden state matrix, effectively facilitating long-range memory. Then, the Selection Mechanism empowers SSMs to acquire content-aware representations. Lastly, Mamba crafts two hardware-aware computation algorithms, Parallel Associative Scan and Memory Recomputation, to enhance training efficiency.

3.1.1 ***HiPPO-based Memory Initialization***. Modeling and learning from sequential data represent foundational challenges in contemporary machine learning, forming the bedrock for various tasks, including language modeling, speech recognition, and video processing. A fundamental component for modeling intricate and long-term temporal dependencies lies in memory, encompassing the ability to store and integrate information from preceding time steps [73]. Similar to RNNs, preserving and forgetting the historical hidden states (i.e., the matrix $\mathbf{A}$) plays a critical role in SSMs to achieve satisfying performances. In previous structured state space sequence models (SSMs), there have been suggestions for special initializations, especially in the case of complex-valued models. These special initializations have proven beneficial in various scenarios, including situations with limited data availability. Similarly, Mamba focuses primarily on the initialization of the hidden state matrix $\mathbf{A}$ to capture complex temporal dependencies. This is accomplished through the utilization of the HiPPO theory [56] with an innovative scaled Legendre measure (LegS), ensuring a comprehensive consideration of the complete historical context rather than a limited sliding window. To be specific, the HiPPO-LegS

assigns uniform weight to all historical data points, which can be expressed as:

$$A_{nk}^{\text{HiPPO}} = -\begin{cases} (2n+1)^{\frac{1}{2}}(2k+1)^{\frac{1}{2}} & \text{if } n > k \\ n+1 & \text{if } n = k \text{ ,} \\ 0 & \text{if } n < k \end{cases} \tag{16}$$

where n is the number of polynomials, and k denotes the particular discrete time steps. Building upon the HiPPO theory, Mamba introduces two simple initialization methods for the complex and real cases, i.e., S4D-Lin and S4D-Real [57], as presented in

$$A_{dn} = -\begin{cases} \frac{1}{2} - ni & \text{S4D-Lin} \\ n+1 & \text{S4D-Real} \end{cases} , \tag{17}$$

where n is the n-th element of A for all input dimensions $d = 1, 2, ..., D$. Given such an initialization, the model can learn long-dependent memory that experiences smaller degradation of newer steps and larger degradation of older steps by compressing and reconstructing the input information signal. According to the formulas, HiPPO-LegS possesses advantageous theoretical properties: it remains consistent across input timescales and offers rapid computation [56]. Additionally, it has bounded gradients and approximation errors, facilitating the parameter learning process.

3.1.2 **Selection Mechanism**. Conventional state space models are unable to produce personalized outputs based on specific model inputs (i.e., the content-aware modeling ability) due to the property of Time Invariance. To provide SSMs with such a capability similar to the attention mechanisms, Mamba designs a time-varying selection mechanism that parameterizes the weight matrices according to model input. Such innovation empowers SSMs to filter out extraneous information while retaining pertinent details indefinitely. Formally, the selection mechanism involves setting the interval Δ, and matrices B, C as functions of the input $x \in \mathbb{R}^{B \times L \times D}$, which can be formulated as:

$$B \rightarrow S^B = W^B x, \tag{18}$$

$$C \rightarrow S^C = W^C x, \tag{19}$$

$$\Delta \rightarrow S^\Delta = \tau_\Delta \cdot \text{BroadCast}_D(W^\Delta x), \tag{20}$$

where $S^B \in \mathbb{R}^{B \times L \times N}$, $S^C \in \mathbb{R}^{B \times L \times N}$, and $S^\Delta \in \mathbb{R}^{B \times L \times D}$ are the selective space matrices that function of the input to achieve content-aware modeling. B, L, D, and N represent the batch size, input length, input feature size, and hidden channel number, respectively. Notably, $W^B \in \mathbb{R}^{N \times D}$, $W^C \in \mathbb{R}^{N \times D}$, and $W^\Delta \in \mathbb{R}^{D \times 1}$ are the selection weights (i.e., linear parameterized projections) for corresponding components, and BroadCast_D means to broadcast the result to all the dimensions $d = 1, 2, .., D$. Subsequently, the selective SSMs undergo discretization using a common statistical technique, Zero-Order Hold (ZOH) [138], as presented in

$$\overline{A} \rightarrow S^{\overline{A}} = \exp(S^\Delta A), \tag{21}$$

$$\overline{B} \rightarrow S^{\overline{B}} = (S^\Delta A)^{-1}(\exp(S^\Delta A) - I) \cdot S^\Delta S^B, \tag{22}$$

where $S^{\overline{A}} \in \mathbb{R}^{B \times L \times D \times N}$ and $S^{\overline{B}} \in \mathbb{R}^{B \times L \times D \times N}$ are the selective state transition matrix and the input matrix, respectively, which become the functions of input x. By doing so, the discrete SSM has changed from time-invariant to time-varying

(i.e., content-aware) as

$$y = \text{SSM}(\mathbf{A}, \mathbf{B}, \mathbf{C})(\mathbf{x}), \tag{23}$$

which generates output $\mathbf{y} \in \mathbf{R}^{B \times L \times D}$ depending on the input $\mathbf{x}$. Note that the time-varying selection mechanism in Mamba has a similar structure to the attention mechanism in Transformer, i.e., both perform operations based on inputs and their projections, which allows Mamba's SSM to achieve a flexible content-aware modeling. Nevertheless, it loses the equivalence to convolutions, which negatively impacts its efficiency.

3.1.3 ***Hardware-aware Computation****.* The selection mechanism is crafted to surpass the limitations of linear time-invariant models. Still, it challenges efficient training: SSMs' convolutional kernels become input-dependent, resulting in the inability to perform parallel computations. To tackle the problem, Mamba utilizes two computation techniques, i.e., *Parallel Associative Scan* (also called Parallel Prefix-Sum) [64] and *Memory Recomputation*. First, the Parallel Associative Scan leverages the property of linear associative computation and the parallelism of modern accelerators (GPU and TPU) to perform the calculation of selective SSMs in a memory-efficient manner. More specifically, the parallel associative scan reduces the computation complexity of model training from $\mathbf{O}(N^2 d)$ to $\mathbf{O}(N/t)$. At its core, the scan revolves around constructing a balanced binary tree on the given input and sweeps it to and from the root. In other words, the parallel associative scan begins by traversing from the leaves to the root (i.e., Sweep-Up), creating partial sums at the internal nodes of the tree. Then, it reverses the traversal, moving from the root back up the tree to construct the whole scan using the partial sums (i.e., Sweep-Down).

On the other hand, Mamba leverages the traditional approach of recomputation to diminish the overall memory demand for training selective SSM layers. In particular, Mamba abstains from storing intermediate states of size (B, L, D, N) during the forward pass of the Parallel Associative Scan to prevent memory expansion. Instead, it recomputes those intermediate states in the backward pass for gradient computation. By doing so, recomputation sidesteps the necessity of reading $O(BLND)$ elements between GPU memory cells. In addition to optimizing the memory needs of the scan operation, Mamba-1 extends its use of recomputation to enhance the efficiency of the entire SSM layer. This optimization encompasses projections, convolutions, and activations, which typically demand significant memory resources but can be rapidly recomputed.

3.2 Mamba-2: State Space Duality

Transformers, which have played a crucial role in the success of deep learning for various areas, have inspired the development of various techniques, such as Parameter-efficient Fine-tuning [95], Catastrophic Forgetting Mitigation [96], and Model Quantization [195], aimed at improving model performance from diverse perspectives. To enable state space models to access and benefit from the valuable techniques initially developed for Transformers, Mamba-2 [29] have introduced a comprehensive framework called Structured State-Space Duality (SSD), which establishes theoretical connections between SSMs and different forms of attention. Formally,

$$y = \text{SSD}(\mathbf{A}, \mathbf{B}, \mathbf{C})(\mathbf{x}) = \mathbf{M}\mathbf{x}, \tag{24}$$

where $\mathbf{M}$ denotes the matrix form of SSMs that uses the sequentially semi-separable representation, and $\mathbf{M}_{ji} = \mathbf{C}_j^{\mathsf{T}} \mathbf{A}_{j:i} \mathbf{B}_i$. Notably, $\mathbf{C}_j$ and $\mathbf{B}_i$ represent the selective space state matrices associated with input tokens $\mathbf{x}_j$ and $\mathbf{x}_i$, respectively. $\mathbf{A}_{j:i}$ denotes the selective matrix of hidden states corresponding to the input tokens ranging from j to i. In essence, SSD demonstrates that both the attention mechanism used by Transformers and the linear time-variant system employed in
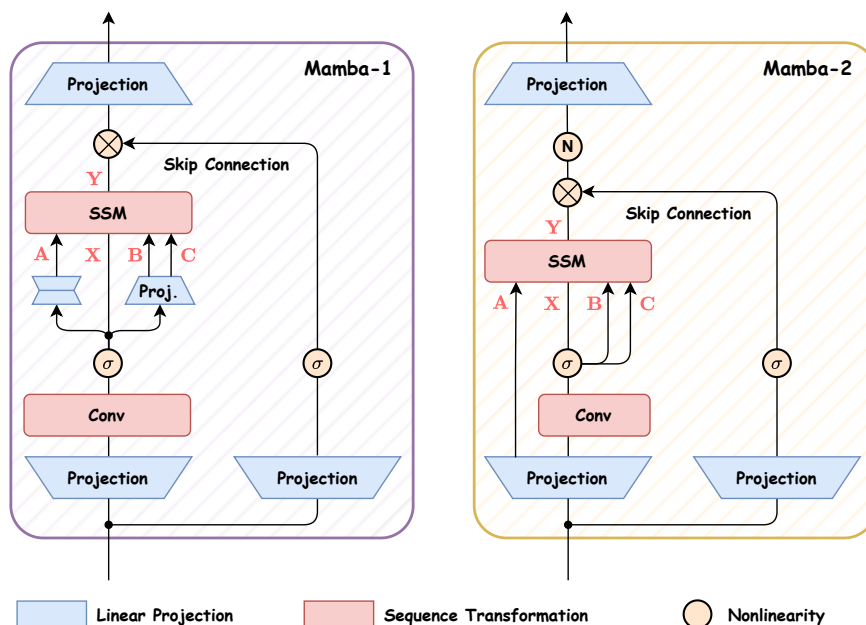
Fig. 4. The block architectures of Mamba-1 and Mamba-2.

SSM can be seen as semi-separable matrix transformations. Furthermore, Dao and Gu [29] also proves that the selective SSM is equivalent to a structured linear attention mechanism implemented with a semi-separable masking matrix.

Based on SSD, Mamba-2 has devised a more hardware-efficient computation through a block decomposition matrix multiplication algorithm. Specifically, by viewing state space models as semi-separable matrices through the matrix transformation, Mamba-2 decomposes the computation into matrix blocks, in which diagonal blocks represent intra-chunk computations. In contrast, the off-diagonal blocks represent inter-chunk computations factored through the SSM's hidden state. This approach enables Mamba-2 to achieve a 2-8× faster training process than Mamba-1's parallel associative scan while remaining competitive with Transformers.

3.3 Mamba Block

In this subsection, we provide a summary of the block design for Mamba-1 and Mamba-2. Figure 4 illustrates the comparison of these two architectures. Mamba-1 is motivated by an SSM-centric point of view where the selective SSM layer is tasked with conducting a map from input sequences $\mathbf{X}$ to $\mathbf{Y}$. In this design, the linear projections of ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$) are applied after the initial linear projection that creates $\mathbf{X}$. The input tokens and state matrices are then passed through the selective SSM cell, utilizing the parallel associative scan, to produce the output $\mathbf{Y}$. After that, Mamba-1 employs a skip connection to encourage feature reuse and alleviate the degradation problem often occurring during the model training process. Finally, the Mamba model is constructed by stacking this block interleaved with standard normalization and residual connections.

As for Mamba-2, it introduces the SSD layer aiming to create a map from [$\mathbf{X}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$] to $\mathbf{Y}$. This is achieved by simultaneously processing [$\mathbf{X}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$] with a single projection at the beginning of the block, similar to how standard attention architectures generate the $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ projections in parallel. In other words, the Mamba-2 block simplifies

the Mamba-1 block by removing sequential linear projections. This enables faster computation of the SSD structure compared to the parallel selective scanning in Mamba-1. Additionally, a normalization layer is added after the skip connection, aiming to improve training stability.

4 Advancements in Mamba Models

State Space Models and Mamba have been recently explored and become one promising alternative as the foundational model backbone. While Mamba demonstrates proficiency in natural language processing, it still encounters challenges, such as memory loss, generalization to diverse tasks, and inferior capability to capture complex patterns to Transformer-based language models. To overcome these challenges, plenty of efforts have been made to improve the Mamba architecture. Existing research studies primarily concentrate on modifying the *block design*, *scanning mode*, and *memory management* aspects. This section will introduce several vital techniques from these three aspects, and a summary of related studies is presented in Table 1.

Table 1. Summary of Existing Studies on Improving the Mamba Model.

Modules	Methods	Classes	Representative References
Block	Integration	Transformer	[65, 111, 141, 142, 200]
		Convolutional Neural Network (CNN)	[53, 107, 107, 161, 187, 204, 213, 214]
		Graph Neural Network (GNN)	[9, 103, 115, 178, 203]
		Recurrent Neural Network (RNN)	[32, 76, 170]
		Spiking Neural Network (SNN)	[8, 106]
	Substitution	U-Net	[30, 71, 84, 110, 118, 125, 152, 153, 157, 163, 179, 180]
		Diffusion Models	[47, 49, 136, 187, 209]
		Others	[19, 101]
	Modification	Mix-of-Expert	[5, 111]
		K-way/Parallel Structure	[79, 113, 177, 193, 235]
		Register	[179, 202]
Scan	Flatten	Bidirectional Scan	[87, 101, 105, 234]
		Sweeping Scan	[121, 188, 214]
		Continuous Scan	[66, 75, 201]
		Efficient Scan	[139, 197]
	Stereo	Hierarchical Scan	[12, 24, 26, 63, 162, 180]
		Spatiotemporal Scan	[21, 102, 205, 207]
		Hybrid Scan	[10, 30, 33, 53, 68, 163]
Memory		Initialization [36], Compression [123, 134], Connection [67, 149]	
Others		Autoregressive Pretraining [150], Explainability [83]	

4.1 Block Design

The design and structure of the Mamba block have a significant impact on the overall performance of Mamba models, making it an emerging research focus. As illustrated in Figure 5, based on different approaches to constructing new Mamba blocks, existing research can be categorized into three categories: a) **Integration** methods aim to integrate the Mamba block with other well-known models, so as to strike a balance between effectiveness and efficiency; b) **Substitution** methods attempt to utilize Mamba block as a substitution for main layers in advanced model frameworks; and c) **Modification** methods focus on modifying the components within the classical Mamba block. Accordingly, we will present a detailed review of these methods in the following subsections.
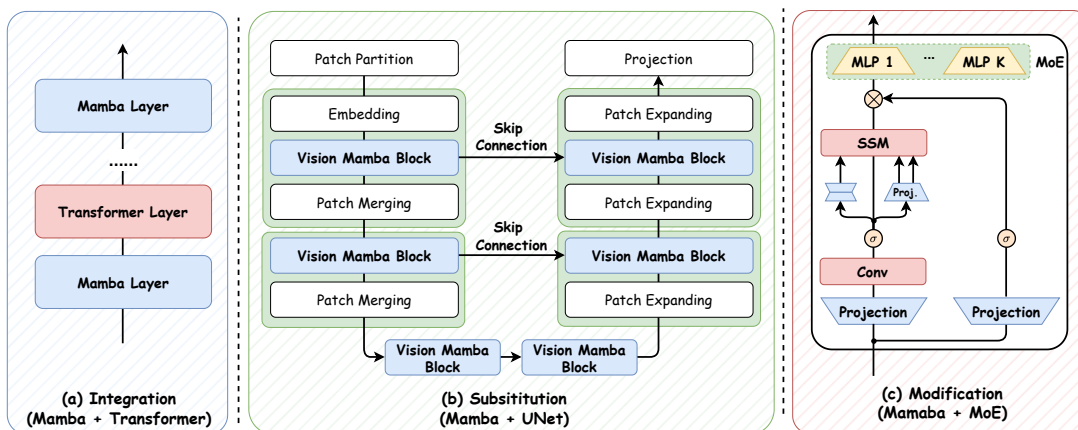
Fig. 5. Representative examples of improved Mamba models based on the perspective of block design: (a). Integration methods combine orthogonal architectural designs (e.g., Transformer) with Mamba, leading to enhanced model performance and increased throughput, all while upholding a manageable memory footprint [141, 200]; (b). Substitution methods enhance the learning capabilities of standard learning frameworks (e.g., UNet) by integrating the Mamba block as a substitute for their primary layers [110, 214]; (c). Modification methods employ cutting-edge techniques, such as Mix-of-Expert (MoE), to refine the Mamba block [111].

4.1.1 ***Integration***. Given Mamba's exceptional ability to capture long-term dynamics, it has been extensively integrated with other models, leveraging their own strengths to deliver a robust framework tailored to specific scenarios. The integration specifically encompasses advanced models like Transformers, Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), Recurrent Neural Networks (RNNs), and Spiking Neural Networks (SNNs). Specific examples are described below.

- **Transformer**-based models have exhibited remarkable performance in numerous tasks, but their quadratic computational complexity still hampers them during inference process [58]. In the pursuit of efficient generation, some researchers have proposed incorporating Mamba blocks with Transformer-based models. For example, Jamba [111] combines blocks of Transformer and Mamba layers to tackle long-content Natural Language Processing tasks, capitalizing on the advantages of both model families. The Attention-Mamba hybrid model demonstrated superior performance compared to the standalone Transformer and Mamba models, achieving better throughput than the vanilla Transformer model. Mambaformer [200] utilizes the hybrid framework to forecast multiple time series, including exchange rates, hourly electricity consumption, and power load, which internally combines Mamba blocks and Transformer layers for long- and short-range dependencies, respectively. Due to the integration of Mamba and Transformer, Mambaformer outperforms Transformer-based predictors in long-short range time series forecasting.
- **CNN**-based methods are constrained by local receptive fields, resulting in suboptimal performance capturing global and long-range semantics [55]. Known for the superior capability of state space models to learn long-range patterns, some studies [107, 187, 204] have explored the potential of utilizing Mamba blocks to enhance CNN-based models, especially in the field of computer vision. For instance, MedMamba [214] and nnMamba [53] showcase how the integration of visual Mamba blocks improves the performance of CNNs in image analysis tasks.
- **GNN** has demonstrated promising potential in capturing neighboring relationships through message-passing mechanisms, where information is propagated over a connection graph through stacked layers. Nonetheless, these

models face a significant limitation known as over-smoothing [20], particularly when attempting to capture high-order adjacency signals. To tackle such a challenge, Mamba has been employed for graph representation learning [103, 115, 178, 203]. For example, Graph Mamba [9] reformulates graph-structured data into sequential tokens in a particular order and leverages a selective SSM layer within the Mamba block to construct a novel Graph Mamba Network (GMN) architecture, which achieves superior graph representation learning capabilities, particularly in the datasets that requires high-order dependencies between nodes.

- **RNN**-based models have yielded outstanding results in capturing temporal dynamics. Nevertheless, RNNs still face significant challenges, including time-consuming recurrent training and limitations in memory capacity for hidden states. Inspired by the emergence of recent Mamba-based architectures, some researchers have developed a fusion of Mamba blocks and RNNs. For instance, VMRNN [170] achieves state-of-the-art performance in spatio-temporal forecasting while minimizing floating-point operations (FLOPs) compared to recurrent-based and recurrent-free methods. It accomplishes this by introducing a novel recurrent unit that combines Mamba blocks with Long Short-Term Memory (LSTM).

- **SNN** has been recently proposed as a promising network architecture inspired by the behavior of biological neurons in the brain: transmitting knowledge between neurons through discrete spikes. One of the key advantages of SNNs lies on their potential for low-power implementation, as they can exploit the sparse and event-driven nature of neural activity. Motivated by the energy-efficient implementation of SNNs and SSMs' superior long-range learning capabilities, pioneering studies have delved into integrating these two methods. For example, SpikeMba [106] combines them to handle confidence bias towards prominent objects and to capture enduring dependencies within video sequences. Through extensive evaluations, the authors claim that integrating these two models improves the effectiveness of temporal video grounding tasks, precisely moment retrieval and highlight detection.

4.1.2 **Substitution**. Inspired by the outstanding capabilities of the selective SSM in efficient computation and long sequence learning, the adoption of Mamba modules to replace critical components in classical modeling frameworks such as U-Net [151] and Diffusion Model [70] has attracted a lot of attention. By introducing the selective SSM layer, these methods achieve long-range learning and efficient computation for their specific tasks. Below, we demonstrate instances of substitution using the Mamba module, specifically for advanced frameworks such as U-Net and Diffusion models.

- **U-Net**. On the one hand, many efforts [110, 163, 179, 180] have been made to synergize U-Net with Mamba's capability in capturing intricate and broad semantics, so as to advance model performance in computer vision tasks. For example, Mamba-UNet [179] utilizes Visual Mamba blocks exclusively to construct a U-Net-like model (i.e., an encoder-decoder model infused with skip connections) for medical image segmentation. Their evaluation demonstrates that Mamba-UNet surpasses several U-Net variations, which can be attributed to the efficacy and efficiency of Mamba blocks in handling long-range patch sequences.

- **Diffusion Model**. On the other hand, some endeavors [47, 49, 136] have been undertaken to build a novel type of diffusion model, Diffusion State Space Model (DiS), which replace the typical backbone (e.g., CNNs, Attentions, U-Nets) with a state space backbone. Given the remarkable efficiency and efficacy of Mamba blocks in accommodating long-range dependencies, DiS is distinguished by generating longer sequences using diffusion models [47]. For example, Oshima et al. [136] propose a Mamba-based diffusion model that substantially decreases memory consumption for long video sequences, while still maintaining competitive performance metrics when compared to Transformer-based

models. Moreover, MD-Dose [49] and P-Mamba [209] construct noise predictors using Mamba blocks in the backward process of diffusion models, ultimately generating specific targets for medical image processing.

- **Others**. Besides the U-Net and Diffusion Models, there are a few substitutions. For example, Res-VMamba [19] adopts Visual Mamba blocks in a residual learning framework for food category classification. Furthermore, SPMamba [101] adopts the TF-GridNet [189], a recently developed time-frequency model, as its base architecture followed by succeeding the Transformer components with bidirectional Mamba blocks. This adaptation enable the model to encompass a wider scope of contextual information efficiently for the task of speech separation.

4.1.3 ***Modification***. Apart from Integrating and Substituting methods that directly employ the Mamba block, some other efforts have been made to modify the Mamba block with the aim of enhancing its performance in different scenarios. For example, Jamba [111] borrows the conception of Mix-of-Experts (MoE) [46, 82] to enable their hybrid (Transformer-Mamba) decoder-only model to be pretrained with far less compute and allow flexible objective-specific configurations. Notably, the Jamba model (56B available parameters, 12B active parameters, 4GB KV cache) requires a 32x smaller KV cache compared to a representative Transformer-based language model, LLaMA-2-7B (6.7B available parameters, 12B active parameters, 128GB KV cache), while providing more extensive available and active parameters. This allows Jamba to swallow a context length of 140K on a single A100 GPU (80GB), seven times the length supported by LLaMA-2-70B. In addition to MoE, some studies propose modifying the SSM layer into a K-way structure, which involves processing model inputs using parallel SSM cells, allowing for capturing information and knowledge from multiple perspectives. For example, Sigma [177] develops a novel Mamba-based visual encoder that handles multimodal inputs by utilizing parallel SSM layers. UltraLight VM-UNet [193] proposes a vision Mamba layer with parallel SSM cells that processes deep features in different channels. To recap, by implementing such modifications (i.e., K-way, MoE), these Mamba-based models gain enhanced learning capabilities, particularly in processing multimodal inputs and fast adapting to multiscale tasks. In addition, a pioneering study, Mamba$^{\circledR}$, has introduced a novel approach that suggests incorporating registers evenly within the visual input tokens before passing the inputs through the SSM layers. This modification aims to enhance the representation of the sequence direction of image patches, thereby enabling the unidirectional inference paradigm of the Mamba block to be applicable to visual tasks. Despite these successes, the exploration of modifying Mamba blocks remains a promising yet under-explored area.

4.2   Scanning Mode

The parallel associative scan operation serves as a crucial component within the Mamba model, which aims to address the computation problem caused by the selection mechanism, accelerate the training process, and reduce memory requirements. It achieves this by leveraging the linear property of time-varying SSMs to design kernel fusion and re-computation at the hardware level. However, Mamba's uni-directional sequence modeling paradigm hinders a comprehensive learning process for various data, such as images and videos. To mitigate this issue, several studies have focused on designing efficient scanning methods to enhance model performance and facilitate the training process of Mamba models. As shown in Figure 6, existing studies that concentrate on developing the scanning mode techniques can be categorized into two classes: 1) **Flatten Scan** approaches process model inputs from a flat perspective of token sequence; and 2) **Stereo Scan** methods scan model inputs across dimensions, channels, or scales.

4.2.1 ***Flattening Scan***. Flattening scan refers to the process of flattening the model input into token sequences and scanning them accordingly from different directions. This type of scanning is commonly employed for both
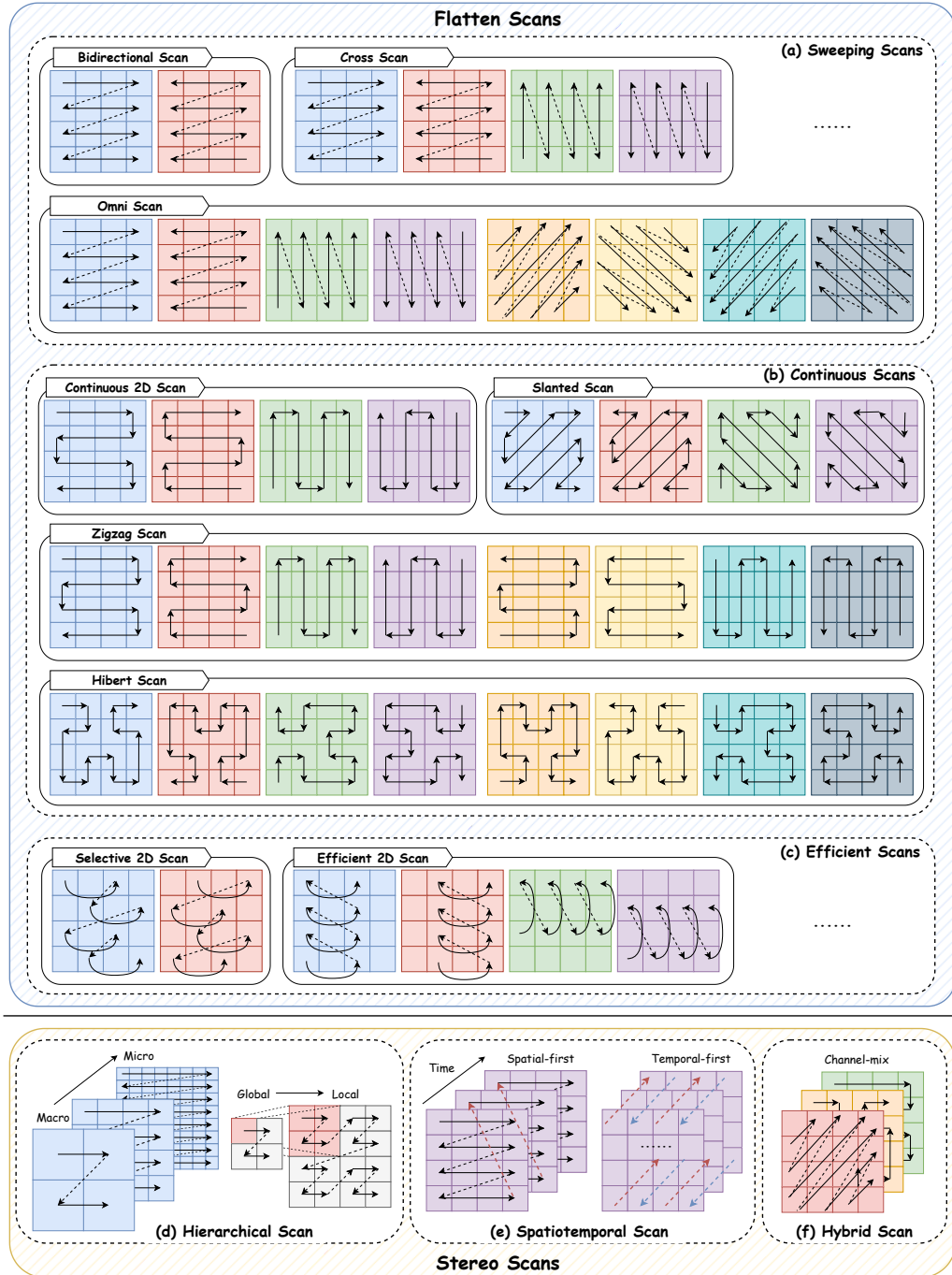
Fig. 6. Recently developed scanning methods in Mamba-based models: Flatten Scans (a-c) involve flattening the model input into token sequences and scanning them accordingly from different directions, and Stereo Scans (d-e) process inputs from additional perspectives, capturing a broader spectrum of knowledge compared to flatten scan methods.

one-dimensional (e.g., time series) and two-dimensional (e.g., image) data. In this section, we further categorize it into four classes, namely **Bidirectional Scan**, **Sweeping Scan**, **Continuous Scan**, and **Efficient Scan**.

- **Bidirectional Scan**. Borrowing the conception of bidirectional recurrent neural networks (Bi-RNNs) [155], Visual Mamba [234] introduces a scanning method for visual data, called Bidirectional Scan (Bi-Scan), which involves processing input tokens using simultaneous forward and backward SSMs, thus enhancing the model capacity for spatially-aware processing. Recently, a number of studies have leveraged the Bi-Scan method to facilitate the learning abilities of their Mamba-based models [105]. For example, DPMamba [87] and SPMamba [101] have both leveraged a pair of dual-path (forward and backward) selective SSMs to model the dependency of speech signals, enabling a bidirectional knowledge processing for speech separation. Such notable successes can be attributed to the effectiveness of Bi-Scan and its ease of deployment.

- **Sweeping Scan.** As illustrated in Figure 6, the Sweeping Scan technique processes the model inputs in a specific direction, analogous to a cleaner meticulously sweeping a floor [188, 214]. For instance, Cross-Scan [121] entails dividing the input image into patches and subsequently flattening it along four distinct paths, which is regarded as a fusion of two bidirectional scans. By adopting these complementary traversal paths, Cross Scan enables each patch in the image to efficiently integrate information from its neighbors in different directions, thereby facilitating the establishment of informative, receptive fields. Omni-Scan [163, 227] incorporates the modeling of image information flows from multiple directions, e.g., 2 (forward and backward) × 4 (left-right, top-bottom, top right-bottom left, top left-bottom right). Such a strategy augments the global modeling capability of contextual information in various directions, enabling the extraction of comprehensive global spatial features.

- **Continuous Scan.** To ensure the continuity of input sequences, Continuous Scan techniques scan the adjacent tokens between columns or rows [66], as shown in Figure 6. For example, in order to better cope with 2D spatial inputs, PlainMamba [201] introduced a continuous scanning approach, known as Continuous Scan, which scans the adjacent tokens between columns (or rows), instead of traveling to the opposite tokens in Cross Scan. Moreover, Hilbert Scan [66] travels a sinuous path based on the Hilbert matrix. Based on their evaluation results, it can be inferred that enhancing the semantic continuity of input tokens leads to superior performance in various visual recognition tasks for Mamba-based models.

- **Efficient Scan**. In contrast to the aforementioned scanning methods, which focus on achieving a more comprehensive input modeling, efficient scanning methods aim to accelerate the training and inference process. Generally, the efficient scan separates the given input into several parts and processes them in parallel, thus reducing computational time. For example, Efficient-2D Scan [139] proceeds images by skipping patches, thus reducing four times of computational demands while preserving global feature maps. Moreover, Gao et al. [51] introduce an effective bi-directional subspace scanning scheme within their Mamba framework. This scheme is designed to capture long-term spatial-angular correspondences efficiently for 4D light field super-resolution tasks. Specifically, it decomposes the patch sequences into two parts and processes them through two bi-directional scanning schemes. By doing so, the scanning method lowers the input length and addresses the long-term memory issues without sacrificing the complete 4D global information.

4.2.2 *Stereo Scan*. By modeling inputs from additional perspectives, stereo-scan methods excel in capturing a broader spectrum of knowledge during the scanning process when compared to flattened scan methods. This enhanced capability allows for a more thorough comprehension of model inputs. To be specific, these methods can be classified into three primary categories: **Hierarchical Scan**, **Spatiotemporal Scan**, and **Hybrid Scan**. The Hierarchical Scan processes

the input from different levels, while the Spatiotemporal Scan considers input patterns from both temporal and spatial perspectives. Additionally, Hybrid Scan combines multiple scanning methods to leverage the benefits of different scan techniques.

- **Hierarchical Scan** methods involve employing different kernel sizes of scanning to capture the semantic knowledge from global to local or from macro to micro perspectives [26, 63, 162, 180]. For example, a Mamba-in-Mamba hierarchical encoder is proposed by [24] for infrared small target detection, combining inner and outer selective SSM blocks. The inner one is specifically tailored to capture the interplay among visual patches for local pattern extraction. Conversely, the outer block is designed to characterize the relationship between visual sentences to capture global features. HiSS [12] divides an input sequence into chunks and models the chunk features hierarchically for continuous sequential prediction. The chunks are first processed by a low-level SSM cell, and the processed features are mapped into an output sequence by a high-level SSM block.
- **Spatiotemporal Scan**. Driven by the prevalence of dynamic systems in the real world, there has been a growing interest in spatiotemporal scanning methods to enhance the performance of Mamba block [205, 207]. For instance, VideoMamba [102] expands the original 2D scan for images into two 3D scans: spatial-first scanning and temporal-first scanning. Combining these two scanning approaches, VideoMamba demonstrates exceptional efficiency in handling long, high-resolution videos. Additionally, ChangeMamba [21] integrates three spatiotemporal scanning mechanisms (sequential modeling, cross modeling, and parallel modeling) to enable contextual information interaction among multi-temporal features for remote sensing change detection.
- **Hybrid Scan**. In the pursuit of comprehensive feature modeling, many efforts have focused on combining the advantages of different scanning methods [30, 33, 53, 163, 229], so-called Hybrid Scan. For example, Mambamixer [10] presents Switch of Scan that dynamically employs a set of image scanning methods, namely Cross-Scan, Zigzag Scan, and Local Scan, to traverse image patches. Mambamixer also introduces a dual selection mechanism to mix information across tokens and channels. By doing so, they show competitive performance with other vision models. Pan-Mamba [68] introduces two scanning methods built upon the Mamba architecture: channel swapping scan and cross-modal scan. By incorporating these two scanning approaches, Pan-Mamba enhances its capabilities in efficient cross-modal information exchange and fusion for image pan-sharpening.

4.3 Memory Management

Like RNNs, the memory of hidden states within state space models effectively stores information from previous steps, thereby playing a crucial role in SSM's overall functionality. While Mamba has introduced the HiPPO-based method for memory initialization [55], challenges still exist in the memory management of the SSM cell, including transferring hidden information between layers and achieving lossless memory compression. To this end, a handful of pioneering studies have proposed different solutions, including memory initialization, compression, and connection. For example, Ezoe and Sato [36] have attempted to refine the initialization process of selective SSMs by using a balanced truncation method during model retraining. Moreover, DGMamba [123] introduces a Hidden State Suppressing method to bolster the domain generalization capabilities of the hidden states within State Space Models. This method works to alleviate the negative effects stemming from these hidden states, thereby narrowing the gap between hidden states across different domains. On a similar note, DenseMamba [67] has put forth a dense connection method to enhance the propagation of hidden information between layers in SSMs. This strategy aims to mitigate memory degradation and preserve detailed information for output generation by selectively integrating hidden states from shallower layers into deeper ones.

5 Adapting Mamba to Diverse Data

The Mamba architecture represents an extension of selective state space models, which possesses fundamental properties of recurrent models that make it well-suited as general foundation models operating on sequences like text, time series, speech, and more. Meanwhile, recent pioneering studies have extended the utilization of the Mamba architecture beyond sequential data, encompassing domains such as images and graphs, as depicted in Figure 7. These studies aim to harness Mamba's remarkable capabilities in capturing long-range dependencies while leveraging its efficiency in learning and inference processes. In this section, we therefore aim to investigate the emerging techniques that adapt Mamba to various types of data.
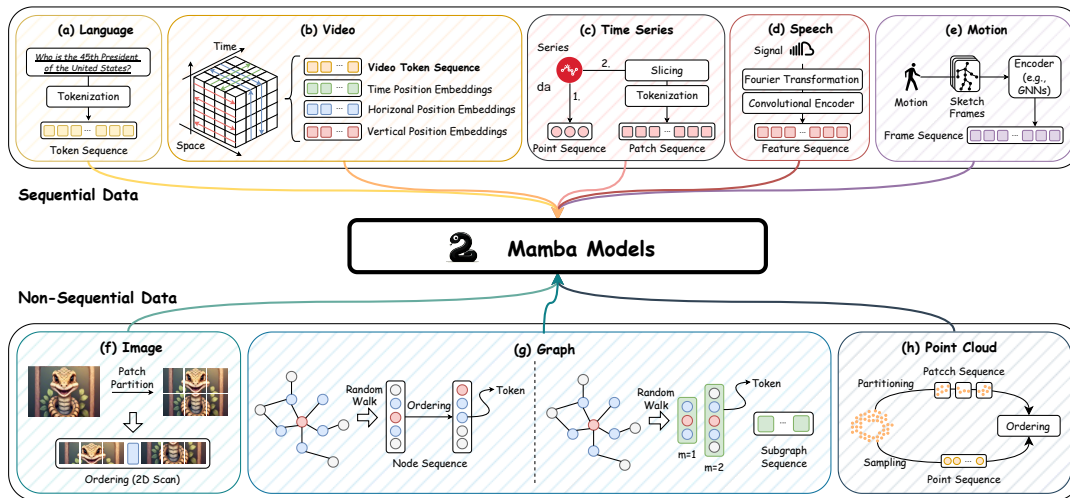

Fig. 7. Representative strategies exist for adapting Mamba to diverse types of data. (a-e) The Mamba architecture, imbued with essential characteristics of recurrent models, serves as an ideal foundational model for handling sequences like language, time series, and speech. (f-h) To address non-sequential data, a common approach involves segmenting or sampling the data into discrete tokens and organizing them into sequences following a defined rule. Additionally, Mamba exhibits the capability to process multimodal data by either concatenating their respective sequences or projections.

Table 2. Summary of Mamba-associated research in different types of data.

Category	Data	Typical Tasks	Representative References
Sequential Data	Language	Long-Context Language Modelling	[5, 10, 12, 58, 67, 111, 134, 143, 160]
	Video	Long Video Generation	[6, 102, 106, 122, 136, 205, 235]
	Time Series	Long-Term Forecasting	[3, 3, 4, 108, 153, 200, 213, 230]
	Speech	Speech Separation	[1, 27, 87, 101]
	Motion	Continuous Human Motion Understanding	[182, 215, 225]
Non-Sequential Data	Image	High-Resolution Medical Vision	[21, 23, 114, 152, 173, 179, 179, 214, 234]
	Graph	Large Graph Learning	[9, 41, 77, 81, 115, 178, 208]
	Point Cloud	Efficient 3D Point Cloud Restoration	[62, 63, 109, 210, 211, 220, 232]
Multimodal Data	Vision-Langue	Visual and Linguistic Awareness	[117, 145, 182, 192, 206]
	Multimodality	Semantic Recognition	[33, 177]

5.1 Sequential Data

Sequential data refers to data gathered or organized in a particular order, where the order of the data points holds significance. To explore the potential of utilizing Mamba as foundation models for tasks concerning sequential data, we provide a comprehensive review presented in the subsequent sections, which cover various sequential data including natural language, video, time-series, speech, and human motion.

5.1.1 **Natural Language***.* As one of the most representative architectures, Mamba performs content-based reasoning while ensuring efficiency, which is considered a promising alternative for the backbone of large language models to address Transformers' computational inefficiency on long sequences. Building upon this insight, many studies have explored the potential of Mamba for various downstream tasks in natural language processing (NLP) [10, 12, 67, 134]. For example, MambaByte [181] is proposed to utilize Mamba on byte sequences, aiming to leverage the advantages of Mamba in capturing long-range dependencies for token-free language models. Their evaluations show that MambaByte avoids the inductive bias of subword tokenization and outperforms state-of-the-art subword Transformers on long-term language modeling tasks. Besides, Jamba [111] and BlackMamba [5] incorporate the concept of Mix-of-Experts (MoE) to enhance Mamba's performance on language processing by integrating the linear-complexity generation from SSMs with the rapid and economical inference capabilities offered by MoE.

5.1.2 **Video***.* The core objective for video understanding and generation lies in learning spatiotemporal representations, which inherently presents two formidable challenges: the large spatiotemporal redundancy within short video clips and the complex spatiotemporal dependencies among long contexts [6, 122]. In the pursuit of addressing both challenges simultaneously, Mamba stands out with its capabilities in distinguishing short-term actions and interpreting long videos [106, 136, 205, 235]. For instance, VideoMamba [102] first projects the input videos into a set of non-overlapping spatiotemporal patches through 3D convolution and then utilizes stacked bidirectional Mamba blocks to encode these patches into vectorized representations for downstream tasks like video understanding and generation.

5.1.3 **Time-Series***.* As typical sequential data, time-series data is ubiquitous in various facets of our lives, including stock market analysis, traffic modeling, and weather forecasting [147, 230]. Motivated by the recent progress on Mamba in modeling long-range sequences, many efforts have been made to investigate its potential for time-series data [3, 4, 108, 200]. For example, TimeMachine [3] harnesses Mamba to capture enduring patterns in multivariate time-series data, ensuring linear-complexity computation and minimal memory footprints for streamlined time-series processing. Moreover, Mambaformer [200] combines selective SSM and Attention layers for the long- and short-term forecasting of weather, traffic flow, and more.

5.1.4 **Speech***.* Speech specifically refers to the vocalized form of human communication that involves vocalized expressions using specific phonetic sounds, words, grammar, and intonation patterns [1]. Recently, in the realm of speech-related tasks, researchers [27] have made significant progress in developing Mamba-based models to tackle the emerging challenges encountered by existing model architectures, such as RNNs and Transformers. For example, SPMamba [101] and DPMamba [87] utilize bidirectional Mamba modules to capture a broader range of contextual information for speech separation, demonstrating a substantial improvement of 13% in model performance and a 566% reduction in computational complexity compared to a Transformer-based baseline when addressing speech separation tasks.

5.1.5 **Motion**. Human motion understanding and generation stand as a significant pursuit in a broad range of practical applications, including computer animation, game development, and robot manipulation. However, semantic actions that occur infrequently within lengthy motion sequences make long-range motion modeling difficult. To address this issue, several studies have proposed the use of Mamba to capture spatiotemporal patterns in motion sequences [182]. For instance, Motion Mamba [225] proposes a hybrid Mamba model, which leverages a hierarchical SSM layer to capture temporal patterns and introduces a bidirectional SSM layer to learn spatial knowledge, preserving motion consistency between frames. Based on the comprehensive experiments, the Mamba-based model outperforms representative diffusion-based methods in human motion generation tasks, achieving a 50% FID improvement and four times faster performance. Additionally, MambaMOS [215] designs a motion-aware state space model that focuses explicitly on capturing variations in motion between consecutive time steps, which further emphasizes the exceptional capabilities of Mamba in achieving high-quality, lengthy sequence motion modeling.

5.2 Non-Sequential Data

Non-sequential data differs from sequential data by not adhering to a specific order. Its data points can be organized or accessed in any sequence without significantly impacting the data's meaning or interpretation [80]. This absence of inherent order presents difficulties for recurrent models such as RNNs and SSMs specifically designed to capture temporal dependencies in data. Surprisingly, Mamba, representing SSMs, has shown outstanding success in efficiently dealing with non-sequential data in recent developments. In this section, we will review relevant studies about how Mamba effectively handles non-sequential data, including images, graphs, and point clouds.

5.2.1 **Image**. As one of the most prevalent modalities, image data forms the foundation of various computer vision applications, e.g., face recognition, medical vision [214], and remote sensing [21, 23]. Drawing inspiration from the success of Mamba in sequence modeling, there exists an intriguing opportunity to transfer this accomplishment from text processing to image analysis. It involves treating an image as a series of patches, potentially paving the way for new avenues of exploration within the realm of computer vision. Thus, plenty of Mamba-based vision models have recently been developed to alleviate heavy computational resources and memory pressures while exhibiting competitive modeling capabilities [114, 152, 179, 179]. For example, Vision Mamba [234] incorporates bidirectional SSM to facilitate global visual semantic modeling and incorporates positional embeddings for location-aware visual comprehension. Not requiring attention mechanisms, Vision Mamba matches the modeling capacity of Vision Transformers while substantially decreasing computation time to subquadratic levels and upholding linear memory complexity. Specifically, it outperforms the state-of-the-art baseline DeiT [173] in terms of speed, being 2.8× faster, and also presents a remarkable reduction of 86.8% in GPU memory usage during batch inference for feature extraction on high-resolution images (1248×1248). Moreover, VMamba [121] introduces 2D Selective Scan (SS2D) that serves as a bridge between 1D array scanning and 2D plane traversal, enabling Mamba to process visual data effectively.

5.2.2 **Graph-structured Data**. Graph modeling has found extensive utility in managing complex structures and relationships, including applications in domains like social networks [41] and molecular interactions [77]. Inspired by the remarkable advent of language foundation models like ChatGPT, numerous researchers have ventured into constructing a graph foundation model that can furnish node and edge embeddings for downstream prediction tasks [28]. Due to the powerful capabilities of Mamba in long-range modeling and high efficiency, several pioneering investigations have embraced the selective State Space Model (SSM) for non-sequential graph data [115]. These studies utilize state space models to encode context through hidden states during recurrent scans, allowing for input flow control, which

resembles attention sparsification on graphs, presenting a data-dependent node selection process within graph modeling contexts [208]. Moreover, Mamba is anticipated to enhance model efficiency during large-graph training tasks.

For example, Graph-Mamba [178] introduces a novel Mamba-based block as a foundational component for graph modeling. This block combines a graph flattening mechanism with the selection mechanism offered by Mamba, transforming sub-graphs into node sequences and facilitating input-dependent context filtering, respectively. In a recent work, Behrouz and Hashemi [9] propose a Graph Mamba Network (GMN), a new graph neural network format based on selective SSMs. The authors reformulate the selective SSM into a graph learning format and provide theoretical justification for the power of the proposed network. By addressing the emerging challenges in crucial steps of graph message passing, GMNs achieve remarkable performance on various aspects, surpassing GNNs and Transformer-based models in multiple benchmark datasets with diverse graph scales. Furthermore, Huang et al. [81] introduce the Graph State Space Convolution (GSSC) as a systematic extension of SSMs tailored for graph-structured data. Specifically, GSSC incorporates distance-based graph convolution kernels into the SSM cell, aiming at enhancing expressive power and capture long-range dependencies. Through assessments conducted on ten benchmark datasets, the study [81] underscores the potential of GSSC as a potent and scalable model for graph machine learning.

5.2.3 **Point Cloud**. Point cloud is a crucial modality in computer vision, with a multitude of practical applications across domains like robotics, autonomous driving, and augmented reality [62]. Unlike image processing and graph learning, the analysis of point cloud presents unique challenges stemming from point clouds' inherent irregularity and sparsity, a 3D non-structured data. To tackle these challenges, notable advancements have been made in deep learning-based approaches, with particular emphasis on Transformer-based models [211]. However, the complexity of attention mechanisms is quadratic, bringing significant computational cost, which is not friendly to low-resource devices. Noted by the recent advance of State Space Models (SSMs) in handling 1D sequences (e.g., language and speech) and 2D data (e.g., image and graph), there have been efforts to extend the application of Mamba to 3D point clouds [210]. In general, these Mamba-based methods for point cloud analysis employ a two-step process [63, 232]. First, the point cloud data is tokenized into discrete tokens using specific scanning methods. Then, Mamba is utilized to capture the underlying patterns within these tokens. For instance, PointMamba [109] proposes a hierarchical scanning strategy to encode local and global information of 3D point cloud and then utilizes plain Mamba as the backbone to extract features from serialized point tokens without incorporating additional complex techniques. Point Cloud Mamba [220] incorporates Mamba as the foundational model backbone to significantly reduce memory usage, demonstrating comparable (or superior) performance compared to Transformer-based counterparts.

5.3 Multimodal Data

Integrating multiple modalities, such as language (sequential data) and images (non-sequential data), offers valuable and complementary information for artificial intelligence perception and scene understanding. Recently, there has been significant research attention on Multimodal Large Language Models (MLLMs) that inherit the advanced capabilities of Large Language Models [192], including powerful language expression and logical reasoning. While Transformers have been the dominant approach in this field, Mamba has emerged as a strong competitor by demonstrating impressive performance in aligning mixed-source data and achieving linear complexity scaling in sequence length, which makes Mamba a promising alternative to Transformers for multimodal learning [117, 206]. For example, Qiao et al. [145] propose VL-Mamba to explore the utilization of Mamba's efficient architecture for solving vision-language tasks, utilizing the pre-trained Mamba model for language understanding and incorporating a connector module to align visual patches with

language tokens. Wang et al. [182] propose Text-controlled Motion Mamba [182], which leverages Mamba to dynamically capture global temporal information based on text queries to enhance human motion understanding. Additionally, Fusion-Mamba [33] and Sigma [177] have tried to fuse complementary information from different modalities such as thermal, depth, and RGB. Fusion-Mamba focuses on improving object detection, while Sigma aims to enhance semantic segmentation.

6 Applications

In this section, we introduce several notable applications of Mamba-based models. To provide a comprehensive overview, we categorize these applications into: *Natural Language Processing*, *Computer Vision*, *Speech Analysis*, *Drug Discovery*, *Recommender Systems*, and *Robotics and Autonomous Systems*.

6.1 Natural Language Processing

In natural language processing domain, recently some Mamba-based models have emerged as alternatives to Transformer-based models for language modeling [5, 14, 67, 111, 176, 226], especially in applications involving extensive contexts such as *Question Answering Systems* and *Text Summarization*.

6.1.1 ***Question Answering Systems****.* Question Answering (QA) involves AI models comprehending, reasoning, and responding using extensive knowledge bases, enabling coherent and contextually rich conversations, widely applied in chatbots and virtual assistants. Incorporating context from previous interactions is crucial for accurately addressing follow-up questions in multi-turn conversations. However, existing models face challenges in inference speed and computational efficiency, particularly in complex reasoning tasks. This leads to significant memory use and computational overhead, which limits scalability and real-time application efficiency. To address these limitations, recent studies have explored Mamba-based models to improve long-term dialogue management in QA Systems [111, 111, 132]. For instance, Mamba-Chat [132] is the first chat language model utilizing a state-space framework. The model maintains and updates its understanding of dialogues by employing state space representations, ensuring context awareness. Jamba [111] strategically alternates between Transformer and Mamba layers, incorporating MoE to enhance model capacity while optimizing parameter utilization. In common sense reasoning and reading comprehension tasks, Jamba achieves performance comparable to larger Llama-2 models but with fewer parameters, demonstrating efficiency and effectiveness. Similarly, DenseMamba [67] introduces a novel method to enrich the propagation of hidden information across layers in SSMs by selectively incorporating hidden states from shallow layers into deeper layers. Compared to traditional Transformer-based models, this preserves crucial fine-grained information for superior performance in question-answering tasks. Overall, integrating Mamba-based models shows promising potential to advance QA systems by improving dialogue management and enhancing performance in complex reasoning tasks.

6.1.2 ***Text Summarization****.* Text summarization aims to condense long texts by preserving essential information. Maintaining coherence and relevance is crucial in this task. Transformer-based models often struggle with long-sequence dependencies, potentially compromising coherence and relevance. In contrast, Mamba-based models leverage robust long-sequence processing capabilities, making them well-suited for processing coherent and context-rich text. Their robust architecture allows them to excel in summarization tasks by accurately capturing and condensing the essence of extensive documents. For instance, LOCOST [14], based on state space models, processes significantly longer sequences than sparse attention models. In long document abstractive summarization, LOCOST achieves performance comparable to the highest-performing sparse transformers of equivalent dimensions while reducing memory usage by up to 50%

during training and 87% during inference. Additionally, SAMBA [149] integrates Mamba with sliding window attention, enabling selective sequence compression into recurrent hidden states while retaining precise memory recall through attention mechanisms. SAMBA achieves a throughput of 3.73 times higher than Transformers when handling input lengths of 128K, showcasing superior performance in tasks requiring long-context summarization.

6.2 Computer Vision

In addition to NLP applications, Mamba-based models have shown potential in the computer vision domain, representative applications like *Disease Diagnosis* and *Motion Recognition and Generation*.

6.2.1 **Disease Diagnosis.** In clinical practice, medical images and videos provide critical insights into the morphology of organs or tissues. Efficient analysis of biomedical objects, such as lesions in large-scale 2D/3D images or videos, significantly enhances disease diagnosis and clinical treatment. However, CNN-based models like UNet face challenges in handling long-range dependencies because of their restricted receptive fields. This challenge is intensified by the typically larger sizes and higher resolution of medical images than natural images. Meanwhile, Transformer-based algorithms are computationally intensive, limiting their practicality in resource-constrained clinical settings. To overcome these limitations, numerous studies have adopted Mamba-based models in real medical environments [110, 127, 152, 187]. For instance, U-Mamba [127] and SegMamba [198] both integrate a hybrid CNN-SSM block, merging the local feature extraction capabilities of convolutional layers with the long-range dependency modeling offered by SSMs. This hybrid approach outperforms the existing models in tasks such as 3D segmentation of abdominal organs in CT and MR images, segmentation of instruments in endoscopy images, and segmentation of cells in microscopy images. Similarly, CMViM [202] addresses challenges in Alzheimer's disease (AD) diagnostic imaging by leveraging masked Vim autoencoders and contrastive learning across modalities, achieving the best performance in AD diagnostic imaging classification. Additionally, ProMamba [196] specializes in polyp segmentation. By incorporating Vision-Mamba architecture and prompt technology, This model achieves higher accuracy and better generalization than previous methods. For dynamic medical object segmentation in videos, Vivim [205] effectively compresses long-term spatiotemporal representations across different scales into sequences using the Temporal Mamba Block. This approach demonstrates enhanced performance and computational efficiency in disease diagnosis, such as ultrasound breast lesions segmentation and polyp segmentation in colonoscopy videos.

6.2.2 **Motion Recognition and Generation.** Motion recognition and generation are critical in motion monitoring [52], computer animation [165], game development [133], and film production [184]. However, transformer-based models encounter challenges related to computational and memory demands, limiting their applicability in resource-constrained environments. Additionally, Transformers and GCNs-based models struggle with effectively capturing long motion sequences and complex spatial-temporal patterns in videos and 4D point clouds. Recent studies have explored the use of Mamba to address these challenges, leveraging its robust performance and lower computational demands [18, 105, 224, 225]. For instance, HARMamba [105] utilizes a bidirectional SSM architecture to process data from wearable sensors, significantly reducing computational load and memory usage while maintaining high accuracy in real-time human motion recognition. Similarly, Simba [18] integrates Mamba within a U-ShiftGCN framework, effectively handling longer sequences and complex spatial-temporal interactions, achieving the best results in skeletal action recognition from videos. Furthermore, Motion Mamba [225] and InfiniMotion [224] are both for motion generation. Specifically, Motion Mamba [225] utilizes hierarchical temporal Mamba blocks for processing temporal data and bidirectional spatial Mamba blocks for handling latent poses, ensuring motion consistency across frames and enhancing

motion generation accuracy within temporal frames. InfiniMotion [224] introduces the Motion Memory Transformer with Bidirectional Mamba Memory, improving the transformer's memory capabilities to efficiently generate continuous, long-duration human motions (up to one hour and 80,000 frames) without overwhelming computational resources.

6.3 Speech Analysis

Speech signals inherently consist of thousands of samples. While this broad temporal context provides rich acoustic features, it also poses significant computational demands. To process speech signals effectively, several Mamba-based models have been successfully employed in diverse speech applications, notably in *Speech Separation and Tagging* and *Speech Enhancement*.

6.3.1 **Speech Separation and Tagging.** Speech separation involves isolating individual speech signals from a multi-speaker environment. It is critical for enhancing the intelligibility and quality of audio communications. Meanwhile, audio tagging or classification involves mapping audio samples to their corresponding categories. Both tasks depend on capturing short-range and long-range audio sequential patterns. Although transformer-based models have been the leading architecture of these applications, they face significant challenges due to their self-attention mechanisms, which result in quadratic computational and memory costs. Recently, there has been a shift toward employing state space models for speech separation [87, 101] and audio tagging [11, 222]. Specifically, DPMamba [87] utilizes selective state spaces to capture dynamic temporal dependencies in speech signals, encompassing both short-term and long-term forward and backward dependencies. SPMamba [101] integrates the TF-GridNet model, replacing its transformer components with bidirectional Mamba modules. DASS [11] integrates knowledge distillation with state-space models, allowing for tagging sound events in audio files lasting up to 2.5 hours. Meanwhile, MAMCA [222] focuses on Automatic Modulation Classification (AMC) by introducing the selective state-space model as its backbone, effectively addressing both accuracy and efficiency challenges associated with long-sequence AMC. By adopting state-space models, these models demonstrate a qualitative improvement, capturing a broader range of contextual information and enhancing overall effectiveness, thereby proving the superior scalability of SSMs in handling long durations.

6.3.2 **Speech Enhancement.** Speech enhancement (SE) aims to extract clear speech components from distorted signals, producing enhanced signals with improved acoustic characteristics. As a front-end processor, SE is pivotal in numerous speech applications, including assistive hearing technologies [98], speaker recognition [7], and automatic speech recognition [129]. Mobile audio devices face challenges due to limited resources. Recent studies have explored the application of Mamba, leveraging its powerful performance and reduced computational demands in SE tasks [17, 148, 159, 168, 221]. For instance, TRAMBA [168] leverages a hybrid architecture combining Transformers and Mamba to improve speech quality for mobile and wearable platforms, specifically targeting acoustic and bone conduction. It achieves a remarkable tenfold reduction in memory consumption compared to current leading models. Additionally, oSpatialNet-Mamba [148] leverages Mamba for long-term multichannel speech enhancement, achieving outstanding results for static and moving speakers.

6.4 Drug Discovery

Protein design, molecular design, and genomic analysis are pivotal in advancing drug discovery and biotechnology [100, 156]. Leveraging the MAMBA-based model significantly reduces the complexities of modeling long sequences in these domains [61, 140, 140, 154, 154]. Specifically, PTM-Mamba [140] and ProtMamba [158] are protein language models based on the Mamba architecture. PTM-Mamba utilizes bidirectional gated Mamba blocks and structured state space

models, efficiently processing long sequences while reducing computational demands. ProtMamba is designed to be homology-aware yet alignment-free, adept at handling extensive contexts across hundreds of protein sequences. Both models maintain high efficiency and accuracy even with large data sets, providing critical tools for protein design. Meanwhile, generative molecular design aims to simulate molecules with tailored property profiles from a specific distribution. However, current models lack the efficiency required to optimize high-fidelity oracles, directly resulting in low success rates. Saturn [61], applying the Mamba architecture, utilizes its linear complexity and computational efficiency to surpass 22 competing models in drug discovery. Furthermore, understanding genomes is crucial for gaining insights into cellular biology. Challenges in genomic modeling include capturing interactions between distant tokens, considering the impacts of both upstream and downstream regions, and ensuring the complementarity of DNA sequences. Caduceus [154] and MSAMamba [172], both leveraging the Mamba model, excel in addressing these challenges. Caduceus, a DNA foundation model, enhances the Mamba architecture with BiMamba and MambaDNA components for bi-directional modeling and ensuring reverse complement equivariance, significantly outperforming existing models in long-range genomic tasks. Similarly, MSAMamba [172] addresses the limitations of transformer-based models for DNA multiple sequence alignments by implementing a selective scan operation along the sequence dimension. This design extends the training context length of previous methods by eightfold, allowing a more comprehensive analysis of extensive DNA sequences.

6.5 Recommender Systems

Recommender Systems are widely utilized in e-commerce [25, 219, 231] and social networks [38, 39, 43], aiming to capture users' evolving preferences and the interdependencies among their past behaviors [40, 228]. Although transformer-based models have demonstrated effectiveness in recommender systems, they face computational efficiency challenges because of the quadratic complexity of attention mechanisms, especially when dealing with longer sequences of behaviors. Recently, several Mamba-based models have been applied to analyze long-term user behavior for personalized recommendations [16, 115, 166, 185, 203]. For instance, Mamba4Rec [115] pioneers the use of selective state space models for efficient sequential recommendation, enhancing model performance while maintaining inference efficiency. Similarly, RecMamba [203] explores Mamba's effectiveness in lifelong sequential recommendation scenarios (i.e., sequence length $\geq$ 2k), achieving comparable performance to benchmark models while cutting down training time by 70% and reducing memory costs by 80%. Furthermore, EchoMamba4Rec [185] integrates a bidirectional Mamba module with frequency-domain filtering to accurately capture intricate patterns and interdependencies within user interaction data. It demonstrates superior performance over existing models, delivering more precise and personalized recommendations. Additionally, Mamba4KT [16] is designed explicitly for knowledge tracing in intelligent education, leveraging the Mamba model to capture enduring correlations between exercises and student knowledge levels. As educational datasets expand, this method suggests a promising avenue for enhancing prediction accuracy, model efficiency, and resource utilization in knowledge tracing research.

6.6 Robotics and Autonomous Systems

The main goal of robotics and autonomous systems is to develop models capable of comprehending visual environments and performing intricate actions. Multimodal Large Language Models (MLLMs) currently used in robotics face significant challenges in two primary aspects: 1) limited capacity for handling intricate tasks requiring advanced reasoning, and 2) substantial computational expenses with fine-tuning and inference tasks. Due to their advantages in inference speed, memory utilization, and overall efficiency, Mamba-based models are emerging as a promising foundation for

autonomous and intelligent systems [15, 85, 117], promising superior performance and substantial scalability potential. For example, RoboMamba [117] integrates a vision encoder with Mamba to create an end-to-end robotic MLLM. This method aligns visual data with language embeddings by co-training, enhancing the model with visual common sense and robot-specific reasoning while ensuring efficient fine-tuning and inference capabilities. Similarly, Jia et al. [85] introduce MaIL, an imitation learning (IL) policy architecture that uses Mamba as a backbone. MaIL bridges the gap between efficiency and performance in handling sequences of observations. Extensive evaluations of real robot experiments demonstrate that MaIL provides a competitive alternative to traditional, large, and complex Transformer-based IL policies.

7 Challenges and Opportunities

The preceding sections have thoroughly surveyed the latest advanced techniques and varied applications associated with Mamba. However, the studies of Mamba are still in its nascent stages, and there exist considerable challenges and opportunities ahead.

7.1 Mamba-based Foundation Models

By scaling up the model sizes to the billion level over large-scale mixture-of-source corpora, foundation models (FMs) exhibit impressive zero-shot learning capabilities, which has enabled FMs to excel in a wide range of general tasks [13]. As a representative example, recent years have witnessed the booming success of Transformer-based large language models, especially ChatGPT, motivating a growing enthusiasm for exploring foundation models in various domains. Even though Transformers are the main drivers of the success, they suffer from pressing computation and memory efficiency issues [171], which come with the exponentially growing training memory proportional to the attention-based model size and the laboriously auto-regressive decoding during inference. In response to these issues, a promising alternative backbone, i.e., Mamba [29, 55], for foundation models has recently emerged. Mamba offers the content-aware learning capabilities of Transformers while scaling the computation linearly with input length, making it effective in capturing long-range dependencies and enhancing efficiency in both training and inference. Given these advantages, developing Mamba-based foundation models for specific domains holds great potential, which offers an opportunity to address the issues faced by Transformer-based models.

7.2 Hardware-Awareness Computation

Foundation models, characterized by their large sizes and intensive matrix operations like matrix multiplications and convolutions, require cutting-edge hardware such as GPUs and TPUs for high-throughput training and inference. These advanced hardware enable researchers to work with larger datasets and achieve state-of-the-art performance across various domains. Nonetheless, the existing foundation models still fall short of fully exploiting the computational capabilities of the hardware, resulting in limited model efficiency [171]. As a promising alternative for enhancing computation efficiency, Mamba-1 [55] and Mamba-2 [29] put forth hardware-aware computation algorithms, namely the Parallel Associative Scan and the Block-decomposition Matrix Multiplication. These algorithms take into account the inherent characteristics of GPUs and TPUs, including factors such as message transmission between devices, offering a fresh perspective on addressing the computation efficiency problem. Inspired by this, exploring novel hardware-efficient algorithms, such as FlashButterfly [48], to optimize hardware utilization offers a promising avenue for conserving resources and accelerating computation, benefiting not only SSMs but also other architectures like Transformers and RNNs.

7.3 Trustworthy Mamba Models

The development of SSMs is expected to bring significant benefits to various industries, including e-commerce, healthcare, and education. At the same time, being a data-dependent model like many existing architectures, Mamba models could pose severe threats to users and society [130]. These threats arise from several factors like erratic decision-making, privacy concerns, and more. Therefore, ensuring trustworthiness in Mamba models is essential across four critical dimensions [116]: *Safety&Robustness*, *Fairness*, *Explainability*, and *Privacy*.

7.3.1 **Safety&Robustness**. Large foundation models have been proven to be highly vulnerable to adversarial perturbations, which can jeopardize the safety and robustness of these models when deployed in safety-critical applications [45, 135, 190]. Meanwhile, Mamba-based models are not exempt from these vulnerabilities [128]. In the pursuit of being a reliable alternative to Transformer, it is essential to investigate and enhance the safety and robustness of Mamba-based models. To be specific, the model outputs should be robust to small perturbations in their inputs. One potential solution could involve automatically pre-processing prompts before feeding them into Mamba-based models. Additionally, as a representative technique, adversarial machine training [78] can be employed to enhance the safety and robustness of Mamba models.

7.3.2 **Fairness**. Large foundation models, trained on extensive datasets, tend to be unintentionally exposed to the biases and stereotypes present in the extensive training corpus [126], which can can manifest in the generated outputs. For instance, within the domain of LLMss, the biases can lead to discriminatory responses influenced by user profile attributes like gender and age, reinforcing stereotypes and unfairly treating specific user groups [86]. While recent efforts have been made to address the issue of fairness in LLMs, there is still a gap in research regarding the non-discrimination and fairness of Mamba models. Thus, further exploration and study are necessary to bridge this gap.

7.3.3 **Explainability**. Deep learning models have often been criticized for their "black-box" nature, and the explainability of deep learning models has emerged as a popular topic in the research community, which indicates the capacity to comprehend and interpret the decisions or predictions generated by a model [35]. By explaining model predictions, users can make more informed decisions based on the model's outputs. To this end, several techniques have been proposed to provide plausible innate explanations for neural architectures based on attention mechanism [74]. Moreover, researchers have investigated the capabilities of Transformer-based language models to generate natural language descriptions to explain their answers [212]. Although an increasing number of studies have attempted to take full advantage of Mamba, studies on comprehending the functioning of Mamba models are still at an early stage, and further investigation is still needed.

7.3.4 **Privacy**. The protection of privacy builds trust between users and Mamba-based models. When users have confidence that their privacy is respected, they are more likely to engage with the AI systems, share relevant information, and seek assistance without fear of misusing their data. Thus, this trust is vital for the widespread adoption and acceptance of Mamba models. One effective strategy for mitigating privacy risks involves cross-verifying the output of Mamba models and screening sensitive content [93]. Moreover, Federated Learning is poised to bolster privacy during the training of Mamba models, wherein the model is trained on numerous decentralized edge devices or servers housing local data samples, without data exchange. This methodology aids in preserving the localization and privacy of the data. Furthermore, integrating privacy-conscious regularization techniques such as differential privacy constraints during training shows promise in preventing overfitting on sensitive data.

7.4 Applying Emerging Techniques from Transformer to Mamba

The Transformer, being the dominant backbone, has led the AI community to develop numerous unique tools aimed at enhancing the performance of attention-based models. Fortunately, by connecting SSMs and attention, the SSD framework introduced by Mamba-2 [29] allows us to develop a shared vocabulary and library of techniques for Transformer and Mamba. In light of this, an important future direction arises, i.e., to explore how the emerging techniques designed for Transformer-based models can be effectively applied to Mamba-based models.

7.4.1 ***Parameter-efficient Finetuning****.* Large foundation models, scaling up their parameters to billion levels, have witnessed groundbreaking advancement in multiple fields. Nevertheless, their extensive scale and computational requirements present significant challenges when tailoring them for specific downstream tasks. To this end, several parameter-efficient finetuning (PEFT) techniques, including the LoRA [72] and Adapter families [50, 91], have been proposed, which involves minimizing the adjustment of parameters or the need for extensive computational resources during fine-tuning. Drawing inspiration from the recent achievements in employing PEFT for large language models constructed using Transformer layers, the adoption of PEFT for Mamba models has emerged as an intriguing topic, with the goal of broadening their range of applications in downstream tasks. For instance, the deployment of LoRA (Low-Rank Adaptation) is anticipated to facilitate rapid fine-tuning for the SSD models, thus enabling the widespread application of Mamba across various domains. However, the specifics of implementing these PEFT techniques for Mamba-based models are yet to be determined and require further investigation.

7.4.2 ***Catastrophic Forgetting Mitigation****.* Catastrophic forgetting, also known as catastrophic interference, refers to the phenomenon observed in machine learning models where they experience a significant loss in performance on previously learned tasks when trained on new tasks [92]. This issue poses a challenge for foundation models because they need to retain knowledge from pre-training tasks and demonstrate consistent performance across different downstream domains. As a promising architecture of the foundation model, Mamba necessitates a thorough investigation to address catastrophic forgetting issues. RRecent research has suggested resolving this challenge by encapsulating task-specific needs through Reward Maximization and Distribution Matching strategies [96, 97]. Moreover, continual learning methods have also been developed to mitigate catastrophic forgetting in Transformer-based language models [90, 186]. These techniques can also be applied to Mamba models by connecting SSMs and attention but remain under-explored.

7.4.3 ***Retrieval-augmented Generation (RAG)****.* Being among the most sophisticated techniques in AI, RAG can provide dependable and current external knowledge, offering significant utility for a multitude of tasks [31, 99]. Large Language Models have recently showcased groundbreaking language comprehension and generation capabilities, despite encountering inherent limitations like hallucinations and outdated internal knowledge. In light of RAG's potent capacity to offer current and valuable supplementary information, Retrieval-Augmented LLMs have emerged, which leverage extraneous knowledge databases to enhance the generative quality of LLMs [22]. Similarly, RAG can be incorporated with Mamba language models to assist them in producing high-quality outputs, which is a promising future research direction.

8 Conclusion

Mamba, an emerging deep learning architecture, has demonstrated remarkable success across diverse domains, such as language generation, image classification, recommendation, and drug discovery, owing to its powerful modeling capabilities and computational efficiency. Recently, increasing efforts have been made to develop Mamba-based deep

learning models with more powerful representation learning capabilities and lower computation complexity. Given the rapid advancement of Mamba, there arises an urgent demand for a systematic overview. To bridge this gap, in this paper, we provide a comprehensive review of Mamba, focusing on its architecture advancements, data adaptability, and application areas, offering researchers both an in-depth understanding and an overview of the latest developments in Mamba. Additionally, given that Mamba research is still in its nascent stages, we also discuss current limitations and present promising directions for future investigation.

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing* 22, 10 (2014), 1533–1545.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Md Atik Ahamed and Qiang Cheng. 2024. Timemachine: A time series is worth 4 mambas for long-term forecasting. *arXiv preprint arXiv:2403.09898* (2024).

[4] Md Atik Ahamed and Qiang Cheng. 2024. TSCMamba: Mamba Meets Multi-View Learning for Time Series Classification. *arXiv preprint arXiv:2406.04419* (2024).

[5] Quentin Anthony, Yury Tokpanov, Paolo Glorioso, and Beren Millidge. 2024. BlackMamba: Mixture of Experts for State-Space Models. *arXiv preprint arXiv:2402.01771* (2024).

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6836–6846.

[7] Zhongxin Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99.

[8] Malyaban Bal and Abhronil Sengupta. 2024. Rethinking Spiking Neural Networks as State Space Models. *arXiv preprint arXiv:2406.02923* (2024).

[9] Ali Behrouz and Farnoosh Hashemi. 2024. Graph Mamba: Towards Learning on Graphs with State Space Models. *arXiv preprint arXiv:2402.08678* (2024).

[10] Ali Behrouz, Michele Santacatterina, and Ramin Zabih. 2024. Mambamixer: Efficient selective state space models with dual token and channel selection. *arXiv preprint arXiv:2403.19888* (2024).

[11] Saurabhchand Bhati, Yuan Gong, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. DASS: Distilled Audio State Space Models Are Stronger and More Duration-Scalable Learners. *arXiv preprint arXiv:2407.04082* (2024).

[12] Raunaq Bhirangi, Chenyu Wang, Venkatesh Pattabiraman, Carmel Majidi, Abhinav Gupta, Tess Hellebrekers, and Lerrel Pinto. 2024. Hierarchical State Space Models for Continuous Sequence-to-Sequence Modeling. *arXiv preprint arXiv:2402.10211* (2024).

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[14] Florian Le Bronnec, Song Duong, Mathieu Ravaut, Alexandre Allauzen, Nancy F Chen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2024. LOCOST: State-Space Models for Long Document Abstractive Summarization. *arXiv preprint arXiv:2401.17919* (2024).

[15] Jiahang Cao, Qiang Zhang, Ziqing Wang, Jiaxu Wang, Hao Cheng, Yecheng Shao, Wen Zhao, Gang Han, Yijie Guo, and Renjing Xu. 2024. Mamba as Decision Maker: Exploring Multi-scale Sequence Modeling in Offline Reinforcement Learning. *arXiv preprint arXiv:2406.02013* (2024).

[16] Yang Cao and Wei Zhang. 2024. Mamba4KT: An Efficient and Effective Mamba-based Knowledge Tracing Model. *arXiv preprint arXiv:2405.16542* (2024).

[17] Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. 2024. An Investigation of Incorporating Mamba for Speech Enhancement. *arXiv preprint arXiv:2405.06573* (2024).

[18] Soumyabrata Chaudhuri and Saumik Bhattacharya. 2024. Simba: Mamba augmented U-ShiftGCN for Skeletal Action Recognition in Videos. *arXiv preprint arXiv:2404.07645* (2024).

[19] Chi-Sheng Chen, Guan-Ying Chen, Dong Zhou, Di Jiang, and Dai-Shi Chen. 2024. Res-VMamba: Fine-Grained Food Category Visual Classification Using Selective State Space Models with Deep Residual Learning. *arXiv preprint arXiv:2402.15761* (2024).

[20] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 3438–3445.

[21] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. 2024. Changemamba: Remote sensing change detection with spatio-temporal state space model. *arXiv preprint arXiv:2404.03425* (2024).

[22] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.

[23] Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. 2024. Rsmamba: Remote sensing image classification with state space model. *arXiv preprint arXiv:2403.19654* (2024).

[24] Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Jieping Ye, and Nenghai Yu. 2024. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *arXiv preprint arXiv:2403.02148* (2024).

[25] Xiao Chen, Wenqi Fan, Jingfan Chen, Haochen Liu, Zitao Liu, Zhaoxiang Zhang, and Qing Li. 2023. Fairly adaptive negative sampling for recommendations. In *Proceedings of the ACM Web Conference 2023*. 3723–3733.

[26] Ying Chen, Jiajing Xie, Yuxiang Lin, Yuhang Song, Wenxian Yang, and Rongshan Yu. 2024. Survmamba: State space model with multi-grained multi-modal interaction for survival prediction. *arXiv preprint arXiv:2404.08027* (2024).

[27] Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Lv Zhao, and Cunhang Fan. 2024. RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection. *arXiv preprint arXiv:2406.06086* (2024).

[28] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter* 25, 2 (2024), 42–61.

[29] Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *International Conference on Machine Learning (ICML)*.

[30] Rui Deng and Tianpei Gu. 2024. CU-Mamba: Selective State Space Models with Channel Learning for Image Restoration. *arXiv preprint arXiv:2404.11778* (2024).

[31] Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.06211* (2024).

[32] Rares Dolga, Kai Biegun, Jake Cunningham, and David Barber. 2024. RotRNN: Modelling Long Sequences with Rotations. *arXiv preprint arXiv:2407.07239* (2024).

[33] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Xuhui Liu, Juan Zhang, Guodong Guo, and Baochang Zhang. 2024. Fusion-mamba for cross-modality object detection. *arXiv preprint arXiv:2404.09146* (2024).

[34] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5792–5793.

[35] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[36] Haruka Ezoe and Kazuhiro Sato. 2024. Learning method for S4 with Diagonal State Space Layers using Balanced Truncation. *arXiv preprint arXiv:2402.15993* (2024).

[37] Lili Fan, Junhao Wang, Yuanmeng Chang, Yuke Li, Yutong Wang, and Dongpu Cao. 2024. 4D mmWave radar for autonomous driving perception: a comprehensive survey. *IEEE Transactions on Intelligent Vehicles* (2024).

[38] Wenqi Fan, Tyler Derr, Yao Ma, Jianping Wang, Jiliang Tang, and Qing Li. 2019. Deep Adversarial Social Recommendation. In *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*. International Joint Conferences on Artificial Intelligence, 1351–1357.

[39] Wenqi Fan, Qing Li, and Min Cheng. 2018. Deep modeling of social relations for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[40] Wenqi Fan, Xiaorui Liu, Wei Jin, Xiangyu Zhao, Jiliang Tang, and Qing Li. 2022. Graph Trend Filtering Networks for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 112–121.

[41] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference*. 417–426.

[42] Wenqi Fan, Yao Ma, Qing Li, Jianping Wang, Guoyong Cai, Jiliang Tang, and Dawei Yin. 2020. A graph neural network framework for social recommendations. *IEEE Transactions on Knowledge and Data Engineering* 34, 5 (2020), 2033–2047.

[43] Wenqi Fan, Yao Ma, Dawei Yin, Jianping Wang, Jiliang Tang, and Qing Li. 2019. Deep social collaborative filtering. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 305–313.

[44] Wenqi Fan, Shijie Wang, Jiani Huang, Zhikai Chen, Yu Song, Wenzhuo Tang, Haitao Mao, Hui Liu, Xiaorui Liu, Dawei Yin, et al. 2024. Graph machine learning in the era of large language models (llms). *arXiv preprint arXiv:2404.14928* (2024).

[45] Wenqi Fan, Xiangyu Zhao, Qing Li, Tyler Derr, Yao Ma, Hui Liu, Jianping Wang, and Jiliang Tang. 2023. Adversarial Attacks for Black-Box Recommender Systems Via Copying Transferable Cross-Domain User Profiles. *IEEE Transactions on Knowledge and Data Engineering* (2023).

[46] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[47] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. 2024. Scalable Diffusion Models with State Space Backbone. *arXiv preprint arXiv:2402.05608* (2024).

[48] Daniel Y Fu, Elliot L Epstein, Eric Nguyen, Armin W Thomas, Michael Zhang, Tri Dao, Atri Rudra, and Christopher Ré. 2023. Simple hardware-efficient long convolutions for sequence modeling. In *International Conference on Machine Learning*. PMLR, 10373–10391.

[49] Linjie Fu, Xia Li, Xiuding Cai, Yingkai Wang, Xueyao Wang, Yali Shen, and Yu Yao. 2024. MD-Dose: A Diffusion Model based on the Mamba for Radiotherapy Dose Prediction. *arXiv preprint arXiv:2403.08479* (2024).

[50] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[51] Ruisheng Gao, Zeyu Xiao, and Zhiwei Xiong. 2024. Mamba-based Light Field Super-Resolution with Efficient Subspace Scanning. *arXiv preprint arXiv:2406.16083* (2024).

[52] Negar Golestani and Mahta Moghaddam. 2020. Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. *Nature communications* 11, 1 (2020), 1551.

[53] Haifan Gong, Luoyao Kang, Yitao Wang, Xiang Wan, and Haofeng Li. 2024. nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. *arXiv preprint arXiv:2402.03526* (2024).

[54] Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks* (2012), 37–45.

[55] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[56] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems* 33 (2020), 1474–1487.

[57] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems* 35 (2022), 35971–35983.

[58] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).

[59] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems* 34 (2021), 572–585.

[60] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. 2024. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles* (2024).

[61] Jeff Guo and Philippe Schwaller. 2024. Saturn: Sample-efficient Generative Molecular Design using Memory Manipulation. *arXiv preprint arXiv:2405.17066* (2024).

[62] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 12 (2020), 4338–4364.

[63] Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. 2024. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. *arXiv preprint arXiv:2404.14966* (2024).

[64] Mark Harris, Shubhabrata Sengupta, and John D Owens. 2007. Parallel prefix sum (scan) with CUDA. *GPU gems* 3, 39 (2007), 851–876.

[65] Ali Hatamizadeh and Jan Kautz. 2024. MambaVision: A Hybrid Mamba-Transformer Vision Backbone. *arXiv preprint arXiv:2407.08083* (2024).

[66] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. 2024. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2404.06564* (2024).

[67] Wei He, Kai Han, Yehui Tang, Chengcheng Wang, Yujie Yang, Tianyu Guo, and Yunhe Wang. 2024. Densemamba: State space models with dense hidden connection for efficient large language models. *arXiv preprint arXiv:2403.00818* (2024).

[68] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. 2024. Pan-Mamba: Effective pan-sharpening with State Space Model. *arXiv preprint arXiv:2402.12192* (2024).

[69] Michiel Hermans and Benjamin Schrauwen. 2013. Training and analysing deep recurrent neural networks. *Advances in neural information processing systems* 26 (2013).

[70] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[71] Alireza Hosseini, Amirhossein Kazerouni, Saeed Akhavan, Michael Brudno, and Babak Taati. 2024. SUM: Saliency Unification through Mamba for Visual Attention Modeling. *arXiv preprint arXiv:2406.17815* (2024).

[72] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[73] Hao Hu and Guo-Jun Qi. 2017. State-frequency memory recurrent neural networks. In *International Conference on Machine Learning*. PMLR, 1568–1577.

[74] Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. 2023. Seat: stable and explainable attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12907–12915.

[75] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. 2024. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802* (2024).

[76] Chensen Huang, Guibo Zhu, Xuepeng Wang, Yifei Luo, Guojing Ge, Haoran Chen, Dong Yi, and Jinqiao Wang. 2024. Recurrent Context Compression: Efficiently Expanding the Context Window of LLM. *arXiv preprint arXiv:2406.06110* (2024).

[77] Kexin Huang, Cao Xiao, Lucas M Glass, Marinka Zitnik, and Jimeng Sun. 2020. SkipGNN: predicting molecular interactions with skip-graph networks. *Scientific reports* 10, 1 (2020), 21092.

[78] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. 43–58.

[79] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2024. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338* (2024).

[80] Tzu-Kuo Huang and Jeff Schneider. 2011. Learning auto-regressive models from sequence and non-sequence data. *Advances in Neural Information Processing Systems* 24 (2011).

[81] Yinan Huang, Siqi Miao, and Pan Li. 2024. What Can We Learn from State Space Models for Machine Learning on Graphs? *arXiv preprint arXiv:2406.05815* (2024).

[82] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[83] Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. 2024. MambaLRP: Explaining Selective State Space Sequence Models. *arXiv preprint arXiv:2406.07592* (2024).

[84] Zexin Ji, Beiji Zou, Xiaoyan Kui, Pierre Vera, and Su Ruan. 2024. Self-Prior Guided Mamba-UNet Networks for Medical Image Super-Resolution. *arXiv preprint arXiv:2407.05993* (2024).

[85] Xiaogang Jia, Qian Wang, Atalay Donat, Bowen Xing, Ge Li, Hongyi Zhou, Onur Celik, Denis Blessing, Rudolf Lioutikov, and Gerhard Neumann. 2024. MaIL: Improving Imitation Learning with Mamba. *arXiv preprint arXiv:2406.08234* (2024).

[86] Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. 2024. Item-side Fairness of Large Language Model-based Recommendation System. In *Proceedings of the ACM on Web Conference 2024*. 4717–4726.

[87] Xilin Jiang, Cong Han, and Nima Mesgarani. 2024. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. *arXiv preprint arXiv:2403.18257* (2024).

[88] Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence* (2024), 1–9.

[89] RE Kalman. 1960. A new approach to linear filtering and prediction problems. *Trans. ASME, D* 82 (1960), 35–44.

[90] Sudipta Kar, Giuseppe Castellucci, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. 2022. Preventing catastrophic forgetting in continual learning of new natural language tasks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3137–3145.

[91] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems* 34 (2021), 1022–1035.

[92] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[93] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[94] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys* 55, 8 (2022), 1–35.

[95] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[96] Tomasz Korbak, Hady Elsahar, German Kruszewski, and Marc Dymetman. 2022. Controlling conditional language models without catastrophic forgetting. In *International Conference on Machine Learning*. PMLR, 11499–11528.

[97] Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems* 35 (2022), 16203–16220.

[98] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, I Made Wartana, et al. 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. *International Journal of Cognitive Computing in Engineering* 3 (2022), 24–30.

[99] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[100] Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[101] Kai Li and Guo Chen. 2024. SPMamba: State-space model is all you need in speech separation. *arXiv preprint arXiv:2404.02063* (2024).

[102] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977* (2024).

[103] Lincan Li, Hanchen Wang, Wenjie Zhang, and Adelle Coster. 2024. Stg-mamba: Spatial-temporal graph learning via selective state space model. *arXiv preprint arXiv:2403.12418* (2024).

[104] Shiwei Li, Huifeng Guo, Xing Tang, Ruiming Tang, Lu Hou, Ruixuan Li, and Rui Zhang. 2024. Embedding Compression in Recommender Systems: A Survey. *Comput. Surveys* 56, 5 (2024), 1–21.

[105] Shuangjian Li, Tao Zhu, Furong Duan, Liming Chen, Huansheng Ning, and Yaping Wan. 2024. Harmamba: Efficient wearable sensor human activity recognition based on bidirectional selective ssm. *arXiv preprint arXiv:2403.20183* (2024).

[106] Wenrui Li, Xiaopeng Hong, and Xiaopeng Fan. 2024. Spikemba: Multi-modal spiking saliency mamba for temporal video grounding. *arXiv preprint arXiv:2404.01174* (2024).

[107] Zhe Li, Haiwei Pan, Kejia Zhang, Yuhua Wang, and Fengming Yu. 2024. Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion. *arXiv preprint arXiv:2404.08406* (2024).

[108] Aobo Liang, Xingguo Jiang, Yan Sun, and Chang Lu. 2024. Bi-Mamba4TS: Bidirectional Mamba for Time Series Forecasting. *arXiv preprint arXiv:2404.15772* (2024).

[109] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. 2024. PointMamba: A Simple State Space Model for Point Cloud Analysis. *arXiv preprint arXiv:2402.10739* (2024).

[110] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Cehngwei Pan, Yasha Wang, and Liantao Ma. 2024. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246* (2024).

[111] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887* (2024).

[112] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. 2022. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–15.

[113] Baijiong Lin, Weisen Jiang, Pengguang Chen, Yu Zhang, Shu Liu, and Ying-Cong Chen. 2024. MTMamba: Enhancing Multi-Task Dense Scene Understanding by Mamba-Based Decoders. *arXiv preprint arXiv:2407.02228* (2024).

[114] Wei-Tung Lin, Yong-Xiang Lin, Jyun-Wei Chen, and Kai-Lung Hua. 2024. PixMamba: Leveraging State Space Models in a Dual-Level Architecture for Underwater Image Enhancement. *arXiv preprint arXiv:2406.08444* (2024).

[115] Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. 2024. Mamba4Rec: Towards Efficient Sequential Recommendation with Selective State Space Models. *arXiv preprint arXiv:2403.03900* (2024).

[116] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–59.

[117] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. 2024. RoboMamba: Multimodal State Space Model for Efficient Robot Reasoning and Manipulation. *arXiv preprint arXiv:2406.04339* (2024).

[118] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. 2024. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302* (2024).

[119] Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2014. A recursive recurrent neural network for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1491–1500.

[120] Xiao Liu, Chenxu Zhang, and Lei Zhang. 2024. Vision Mamba: A Comprehensive Survey and Taxonomy. *arXiv preprint arXiv:2405.04404* (2024).

[121] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166* (2024).

[122] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3202–3211.

[123] Shaocong Long, Qianyu Zhou, Xiangtai Li, Xuequan Lu, Chenhao Ying, Yuan Luo, Lizhuang Ma, and Shuicheng Yan. 2024. Dgmamba: Domain generalization via generalized state space model. *arXiv preprint arXiv:2404.07794* (2024).

[124] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. 2021. Soft: Softmax-free transformer with linear complexity. *Advances in Neural Information Processing Systems* 34 (2021), 21297–21309.

[125] Chao Ma and Ziyang Wang. 2024. Semi-Mamba-UNet: Pixel-Level Contrastive and Pixel-Level Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation. *arXiv e-prints* (2024), arXiv–2402.

[126] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2024. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[127] Jun Ma, Feifei Li, and Bo Wang. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024).

[128] Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Shahbaz Khan, and Salman Khan. 2024. Towards Evaluating the Robustness of Visual State Space Models. *arXiv preprint arXiv:2406.09407* (2024).

[129] Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80 (2021), 9411–9457.

[130] Joao Marques-Silva and Alexey Ignatiev. 2022. Delivering trustworthy AI through formal XAI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12342–12350.

[131] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[132] Justus Mattern and Konstantin Hohr. 2023. Mamba-Chat. GitHub. https://github.com/havenhq/mamba-chat

[133] Nadia Nasri, Sergio Orts-Escolano, and Miguel Cazorla. 2020. An semg-controlled 3d game for rehabilitation therapies: Real-time time hand gesture recognition using deep learning techniques. *Sensors* 20, 22 (2020), 6451.

[134] Piotr Nawrot, Adrian Łańcucki, Marcin Chochowski, David Tarjan, and Edoardo M Ponti. 2024. Dynamic Memory Compression: Retrofitting LLMs for Accelerated Inference. *arXiv preprint arXiv:2403.09636* (2024).

[135] Liang-bo Ning, Zeyu Dai, Jingran Su, Chao Pan, Luning Wang, Wenqi Fan, and Qing Li. 2024. Interpretation-Empowered Neural Cleanse for Backdoor Attacks. In *Companion Proceedings of the ACM on Web Conference 2024*. 951–954.

[136] Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. 2024. Ssm meets video diffusion models: Efficient video generation with structured state spaces. *arXiv preprint arXiv:2403.07711* (2024).

[137] Badri Narayana Patro and Vijay Srinivas Agneeswaran. 2024. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *arXiv preprint arXiv:2404.16112* (2024).

[138] Georgia Pechlivanidou and Nicholas Karampetakis. 2022. Zero-order hold discretization of general state space systems with input delay. *IMA Journal of Mathematical Control and Information* 39, 2 (2022), 708–730.

[139] Xiaohuan Pei, Tao Huang, and Chang Xu. 2024. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977* (2024).

[140] Zhangzhi Peng, Benjamin Schussheim, and Pranam Chatterjee. 2024. PTM-Mamba: A PTM-Aware Protein Language Model with Bidirectional Gated Mamba Blocks. *bioRxiv* (2024), 2024–02.

[141] Jonathan Pilault, Mahan Fathi, Orhan Firat, Chris Pal, Pierre-Luc Bacon, and Ross Goroshin. 2024. Block-state transformers. *Advances in Neural Information Processing Systems* 36 (2024).

[142] Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso, and André FT Martins. 2024. How Effective are State Space Models for Machine Translation? *arXiv preprint arXiv:2407.05489* (2024).

[143] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. 2023. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*. PMLR, 28043–28078.

[144] Matteo Prata, Giuseppe Masi, Leonardo Berti, Viviana Arrigoni, Andrea Coletta, Irene Cannistraci, Svitlana Vyetrenko, Paola Velardi, and Novella Bartolini. 2024. Lob-based deep learning models for stock price trend prediction: a benchmark study. *Artificial Intelligence Review* 57, 5 (2024), 1–45.

[145] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600* (2024).

[146] Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. 2024. TokenRec: Learning to Tokenize ID for LLM-based Generative Recommendation. *arXiv preprint arXiv:2406.10450* (2024).

[147] Haohao Qu, Haoxuan Kuang, Qiuxuan Wang, Jun Li, and Linlin You. 2024. A physics-informed and attention-based graph learning approach for regional electric vehicle charging demand prediction. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[148] Changsheng Quan and Xiaofei Li. 2024. Multichannel long-term streaming neural speech enhancement for static and moving speakers. *arXiv preprint arXiv:2403.07675* (2024).

[149] Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *arXiv preprint arXiv:2406.07522* (2024).

[150] Sucheng Ren, Xianhang Li, Haoqin Tu, Feng Wang, Fangxun Shu, Lei Zhang, Jieru Mei, Linjie Yang, Peng Wang, Heng Wang, et al. 2024. Autoregressive Pretraining with Mamba in Vision. *arXiv preprint arXiv:2406.07537* (2024).

[151] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[152] Jiacheng Ruan and Suncheng Xiang. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491* (2024).

[153] Kazi Shahriar Sanjid, Md Tanzim Hossain, Md Shakib Shahariar Junayed, and Dr Mohammad Monir Uddin. 2024. Integrating mamba sequence model and hierarchical upsampling network for accurate semantic segmentation of multiple sclerosis legion. *arXiv preprint arXiv:2403.17432* (2024).

[154] Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. 2024. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234* (2024).

[155] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[156] Duncan E Scott, Andrew R Bayly, Chris Abell, and John Skidmore. 2016. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery* 15, 8 (2016), 533–550.

[157] Mohammad Shahab Sepehri, Zalan Fabian, and Mahdi Soltanolkotabi. 2024. Serpent: Scalable and Efficient Image Restoration via Multi-scale Structured State Space Models. *arXiv preprint arXiv:2403.17902* (2024).

[158] Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. 2024. ProtMamba: a homology-aware but alignment-free protein state space model. *bioRxiv* (2024), 2024–05.

[159] Siavash Shams, Sukru Samet Dindar, Xilin Jiang, and Nima Mesgarani. 2024. Ssamba: Self-supervised audio representation learning with mamba state space model. *arXiv preprint arXiv:2405.11831* (2024).

[160] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. 2021. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 3531–3539.

[161] Jiamu Sheng, Jingyi Zhou, Jiong Wang, Peng Ye, and Jiayuan Fan. 2024. DualMamba: A Lightweight Spectral-Spatial Mamba-Convolution Network for Hyperspectral Image Classification. *arXiv preprint arXiv:2406.07050* (2024).

[162] Yuheng Shi, Minjing Dong, and Chang Xu. 2024. Multi-Scale VMamba: Hierarchy in Hierarchy Visual State Space Model. *arXiv preprint arXiv:2405.14174* (2024).

[163] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. 2024. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423* (2024).

[164] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4733–4743.

[165] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13653–13662.

[166] Jinzhao Su and Zhenhua Huang. 2024. MLSA4Rec: Mamba Combined with Low-Rank Decomposed Self-Attention for Sequential Recommendation. *arXiv preprint arXiv:2407.13135* (2024).

[167] Jinsong Su, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu. 2017. Lattice-based recurrent neural network encoders for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[168] Yueyuan Sui, Minghui Zhao, Junxi Xia, Xiaofan Jiang, and Stephen Xia. 2024. TRAMBA: A Hybrid Transformer and Mamba Architecture for Practical Audio and Bone Conduction Speech Super Resolution and Enhancement on Mobile and Wearable Platforms. *arXiv preprint arXiv:2405.01242* (2024).

[169] Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 1017–1024.

[170] Yujin Tang, Peijie Dong, Zhenheng Tang, Xiaowen Chu, and Junwei Liang. 2024. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. *arXiv preprint arXiv:2403.16536* (2024).

[171] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *Comput. Surveys* 55, 6 (2022), 1–28.

[172] Vishrut Thoutam and Dina Ellsworth. [n. d.]. MSAMamba: Adapting Subquadratic Models To Long-Context DNA MSA Analysis. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

[173] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*. PMLR, 10347–10357.

[174] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[175] Jean-Philippe Vert. 2023. How will generative AI disrupt data science in drug discovery? *Nature Biotechnology* 41, 6 (2023), 750–751.

[176] Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. 2024. An Empirical Study of Mamba-based Language Models. *arXiv preprint arXiv:2406.07887* (2024).

[177] Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. 2024. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256* (2024).

[178] Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. 2024. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789* (2024).

[179] Feng Wang, Jiahao Wang, Sucheng Ren, Guoyizhe Wei, Jieru Mei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. 2024. Mamba-R: Vision Mamba ALSO Needs Registers. *arXiv preprint arXiv:2405.14858* (2024).

[180] Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. 2024. Large window-based mamba unet for medical image segmentation: Beyond convolution and self-attention. *arXiv preprint arXiv:2403.07332* (2024).

[181] Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. 2024. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660* (2024).

[182] Xinghan Wang, Zixi Kang, and Yadong Mu. 2024. Text-controlled Motion Mamba: Text-Instructed Temporal Grounding of Human Motion. *arXiv preprint arXiv:2404.11375* (2024).

[183] Xiao Wang, Shiao Wang, Yuhe Ding, Yuehang Li, Wentao Wu, Yao Rong, Weizhe Kong, Ju Huang, Shihao Li, Haoxiang Yang, et al. 2024. State space model for new-generation network alternative to transformers: A survey. *arXiv preprint arXiv:2404.09516* (2024).

[184] Yuwei Wang et al. 2023. 3D dynamic image modeling based on machine learning in film and television animation. *Journal of Multimedia Information System* 10, 1 (2023), 69–78.

[185] Yuda Wang, Xuxin He, and Shengxin Zhu. 2024. EchoMamba4Rec: Harmonizing Bidirectional State Space Models with Spectral Filtering for Advanced Sequential Recommendation. *arXiv preprint arXiv:2406.02638* (2024).

[186] Zhen Wang, Liu Liu, Yiqun Duan, Yajing Kong, and Dacheng Tao. 2022. Continual learning with lifelong vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 171–181.

[187] Ziyang Wang and Chao Ma. 2024. Weak-Mamba-UNet: Visual Mamba Makes CNN and ViT Work Better for Scribble-based Medical Image Segmentation. *arXiv preprint arXiv:2402.10887* (2024).

[188] Ziyang Wang, Jian-Qing Zheng, Chao Ma, and Tao Guo. 2024. Vmambamorph: a visual mamba-based framework with cross-scan module for deformable 3d image registration. *arXiv preprint arXiv:2404.05105* (2024).

[189] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. 2023. TF-GridNet: Making time-frequency domain models great again for monaural speaker separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[190] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems* 36 (2024).

[191] Christopher Williams, Fabian Falck, George Deligiannidis, Chris C Holmes, Arnaud Doucet, and Saifuddin Syed. 2024. A Unified Framework for U-Net Design and Analysis. *Advances in Neural Information Processing Systems* 36 (2024).

[192] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2247–2256.

[193] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. 2024. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035* (2024).

[194] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.

[195] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*. PMLR, 38087–38099.

[196] Jianhao Xie, Ruofan Liao, Ziang Zhang, Sida Yi, Yuesheng Zhu, and Guibo Luo. 2024. ProMamba: Prompt-Mamba for polyp segmentation. *arXiv preprint arXiv:2403.13660* (2024).

[197] Xinyu Xie, Yawen Cui, Chio-In Ieong, Tao Tan, Xiaozhi Zhang, Xubin Zheng, and Zitong Yu. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *arXiv preprint arXiv:2404.09498* (2024).

[198] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560* (2024).

[199] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. 2024. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861* (2024).

[200] Xiongxiao Xu, Yueqing Liang, Baixiang Huang, Zhiling Lan, and Kai Shu. 2024. Integrating Mamba and Transformer for Long-Short Range Time Series Forecasting. *arXiv preprint arXiv:2404.14757* (2024).

[201] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. 2024. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695* (2024).

[202] Guangqian Yang, Kangrui Du, Zhihan Yang, Ye Du, Yongping Zheng, and Shujun Wang. 2024. CMViM: Contrastive Masked Vim Autoencoder for 3D Multi-modal Representation Learning for AD classification. *arXiv preprint arXiv:2403.16520* (2024).

[203] Jiyuan Yang, Yuanzi Li, Jingyu Zhao, Hanbing Wang, Muyang Ma, Jun Ma, Zhaochun Ren, Mengqi Zhang, Xin Xin, Zhumin Chen, et al. 2024. Uncovering Selective State Space Model's Capabilities in Lifelong Sequential Recommendation. *arXiv preprint arXiv:2403.16371* (2024).

[204] Judy X Yang, Jun Zhou, Jing Wang, Hui Tian, and Alan Wee Chung Liew. 2024. Hsimamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification. *arXiv preprint arXiv:2404.00272* (2024).

[205] Yijun Yang, Zhaohu Xing, and Lei Zhu. 2024. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168* (2024).

[206] Zhe Yang, Wenrui Li, and Guanghui Cheng. 2024. SHMamba: Structured Hyperbolic State Space Model for Audio-Visual Question Answering. *arXiv preprint arXiv:2406.09833* (2024).

[207] Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. 2024. Spectralmamba: Efficient mamba for hyperspectral image classification. *arXiv preprint arXiv:2404.08489* (2024).

[208] Yang Ye and Shihao Ji. 2021. Sparse graph attention networks. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2021), 905–916.

[209] Zi Ye and Tianxiang Chen. 2024. P-Mamba: Marrying Perona Malik Diffusion with Mamba for Efficient Pediatric Echocardiographic Left Ventricular Segmentation. *arXiv preprint arXiv:2402.08506* (2024).

[210] Xuanyu Yi, Zike Wu, Qiuhong Shen, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Shuicheng Yan, Xinchao Wang, and Hanwang Zhang. 2024. MVGamba: Unify 3D Content Generation as State Space Sequence Modeling. *arXiv preprint arXiv:2406.06367* (2024).

[211] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19313–19322.

[212] Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024*. 1963–1974.

[213] Doncheng Yuan, Jianzhe Xue, Jinshan Su, Wenchao Xu, and Haibo Zhou. 2024. ST-Mamba: Spatial-Temporal Mamba for Traffic Flow Estimation Recovery using Limited Data. *arXiv preprint arXiv:2407.08558* (2024).

[214] Yubiao Yue and Zhenzhang Li. 2024. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849* (2024).

[215] Kang Zeng, Hao Shi, Jiacheng Lin, Siyu Li, Jintao Cheng, Kaiwei Wang, Zhiyong Li, and Kailun Yang. 2024. MambaMOS: LiDAR-based 3D Moving Object Segmentation with Motion-aware State Space Model. *arXiv preprint arXiv:2404.12794* (2024).

[216] Cheng Zhang, Nilam Nur Amir Sjarif, and Roslina Ibrahim. 2024. Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 14, 1 (2024), e1519.

[217] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *Comput. Surveys* 56, 3 (2023), 1–37.

[218] Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, and Zi Ye. 2024. A survey on visual mamba. *arXiv preprint arXiv:2404.15956* (2024).

[219] Jiahao Zhang, Rui Xue, Wenqi Fan, Xin Xu, Qing Li, Jian Pei, and Xiaorui Liu. 2024. Linear-Time Graph Neural Networks for Scalable Recommendations. In *Proceedings of the ACM on Web Conference 2024*. 3533–3544.

[220] Tao Zhang, Xiangtai Li, Haobo Yuan, Shunping Ji, and Shuicheng Yan. 2024. Point Could Mamba: Point Cloud Learning via State Space Model. *arXiv preprint arXiv:2403.00762* (2024).

[221] Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. 2024. Mamba in Speech: Towards an Alternative to Self-Attention. *arXiv preprint arXiv:2405.12609* (2024).

[222] Yezhuo Zhang, Zinan Zhou, Yichao Cao, Guangyu Li, and Xuanpeng Li. 2024. MAMCA–Optimal on Accuracy and Efficiency for Automatic Modulation Classification with Extended Signal Length. *arXiv preprint arXiv:2405.11263* (2024).

[223] Zheng Zhang and Kil To Chong. 2007. Comparison between first-order hold with zero-order hold in discretization of input-delay nonlinear systems. In *2007 International Conference on Control, Automation and Systems*. IEEE, 2892–2896.

[224] Zeyu Zhang, Akide Liu, Qi Chen, Feng Chen, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. 2024. InfiniMotion: Mamba Boosts Memory in Transformer for Arbitrary Long Motion Generation. *arXiv preprint arXiv:2407.10061* (2024).

[225] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. 2024. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487* (2024).

[226] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520* (2024).

[227] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. 2024. Rs-mamba for large remote sensing image dense prediction. *arXiv preprint arXiv:2404.02668* (2024).

[228] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering* (2024).

[229] Zou Zhen, Yu Hu, and Zhao Feng. 2024. Freqmamba: Viewing mamba from a frequency perspective for image deraining. *arXiv preprint arXiv:2404.09476* (2024).

[230] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[231] Meizi Zhou, Zhuoye Ding, Jiliang Tang, and Dawei Yin. 2018. Micro behaviors: A new perspective in e-commerce recommender systems. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 727–735.

[232] Qingyuan Zhou, Weidong Yang, Ben Fei, Jingyi Xu, Rui Zhang, Keyi Liu, Yeqi Luo, and Ying He. 2024. 3dmambaipf: A state space model for iterative point cloud filtering via differentiable rendering. *arXiv preprint arXiv:2404.05522* (2024).

[233] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. 2021. Long-short transformer: Efficient transformers for language and vision. *Advances in neural information processing systems* 34 (2021), 17723–17736.

[234] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).

[235] Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. 2024. Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. *arXiv preprint arXiv:2404.06483* (2024).