Mamba is a type of structured State Space Models(SSM), whose core state space are defined by Equations 1.

$$\begin{aligned} h(t) &= Ah(t-1) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{1}$$

where A and C represent the parameter matrices for compressing historical information, while B and D are the weight matrices for the current input (with A, B, C and D being learnable parameters). $h(t) \in R^n$ denotes the hidden state vector at time t, $x(t) \in R^m$ is the input vector at time t, and $y(t) \in R^p$ represents the output vector at time t. This differential equation models the evolution of the system's state $h(t-1)$ as a linear combination of the previous state and the current input $x(t)$.

Mamba consists of three main components: selective processing of input information, hardware-aware algorithms with parallel scanning method, and a simplified SSM a rchitecture. Although effective, Transformer attention's reliance on key–value caches incurs high costs. Mamba mitigates this by using a parameterized SSM to compress states, filtering irrelevant details and preserving key i nformation. With input-dependent parameters, Mamba uses a parallel scan for efficient, synchronous sequence computation. Furthermore, Flash Attention technology is used to limit the number of data transfers from DRAM to SRAM, allowing bulk data to be written to DRAM in batches and reducing the frequency of read-write cycles. By combining the SSM architecture with the Gated Multilayer Perceptron (MLP) used in modern neural networks like Transformer, the Mamba block is formed (as shown in Supplemental Fig.S4). Stacking the block with residuals and normalization yields Mamba architecture. Mamba, originally for Natural Language Processing(NLP), evolved into a linear-complexity method, but its application in vision is challenging due to the multi-dimensional nature of visual data. Influenced by ViT, Liu et al. Zhu et al. (2024) adapted the Mamba model from language modeling to the vision domain, proposing Vision Mamba. To tackle unidirectional modeling and the lack of position awareness, the Vim block was introduced. It divides the image into patches, which are linearly mapped to a high-dimensional space for processing. Additionally, Positional embeddings and additional markers are used to indicate that the entire image block undergoes bidirectional SSM for global visual context modeling of data dependencies. The entire process can be mathematically described by Equations 2. This model retains the benefits of ViT serialization and modality-agnostic modeling, while addressing the computational complexity of Transformer. The Vision Mamba model is especially suited for high-resolution or long data sequences.

$$\begin{aligned} T_0 &= [t_{cls}; t_p^1 W; t_p^2 W ... t_p^J W] + E_{pos} \\ T_l &= \mathrm{Vim}(T_{l-1}) + T_{l+1} \\ f &= \mathrm{Norm}(T_0^L) \\ p &= \mathrm{MAP}(f) \end{aligned} \tag{2}$$

where T_0 is the initial input sequence and is the sequence processed through Vim. W is the linear projection matrix, Norm refers to the normalization layer, MLP stands for Multi-Layer Perceptron, and p is the final prediction.