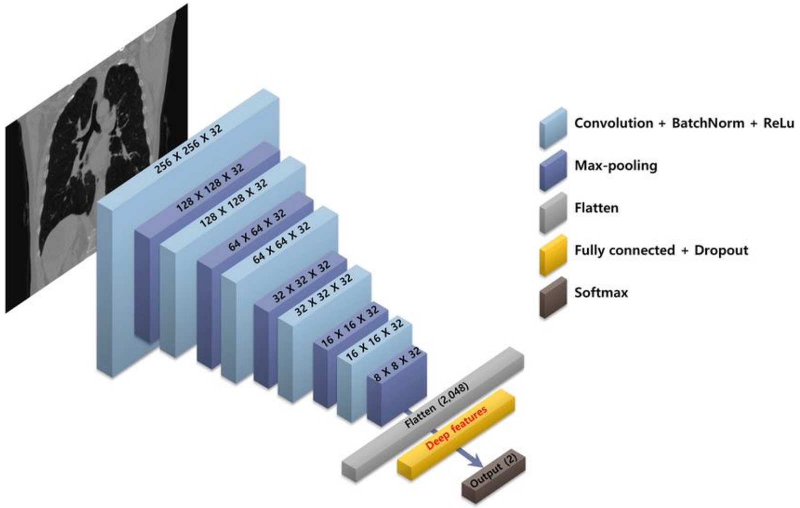
# OuroMamba: A Data-Free Quantization Framework for Vision Mamba Models

**ICCV 2025**

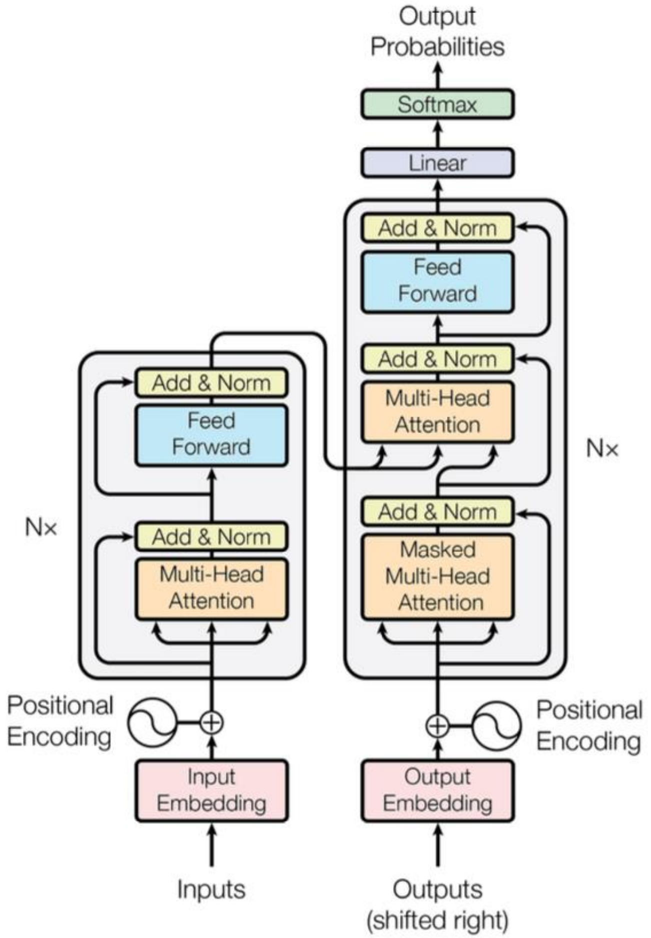
**Akshat Ramachandran**[1*], Mingyu Lee[1*], Huan Xu[1] , Souvik Kundu[2], Tushar Krishna[1]
*Equal Contribution

[1]Georgia Institute of Technology [2]Intel Labs
Contact: akshat.r@gatech.edu

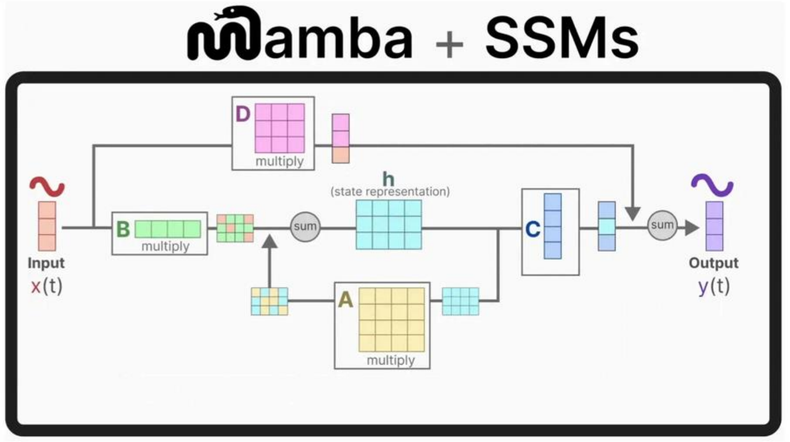
# Diverse Space of AI Models




Convolutional Neural Networks

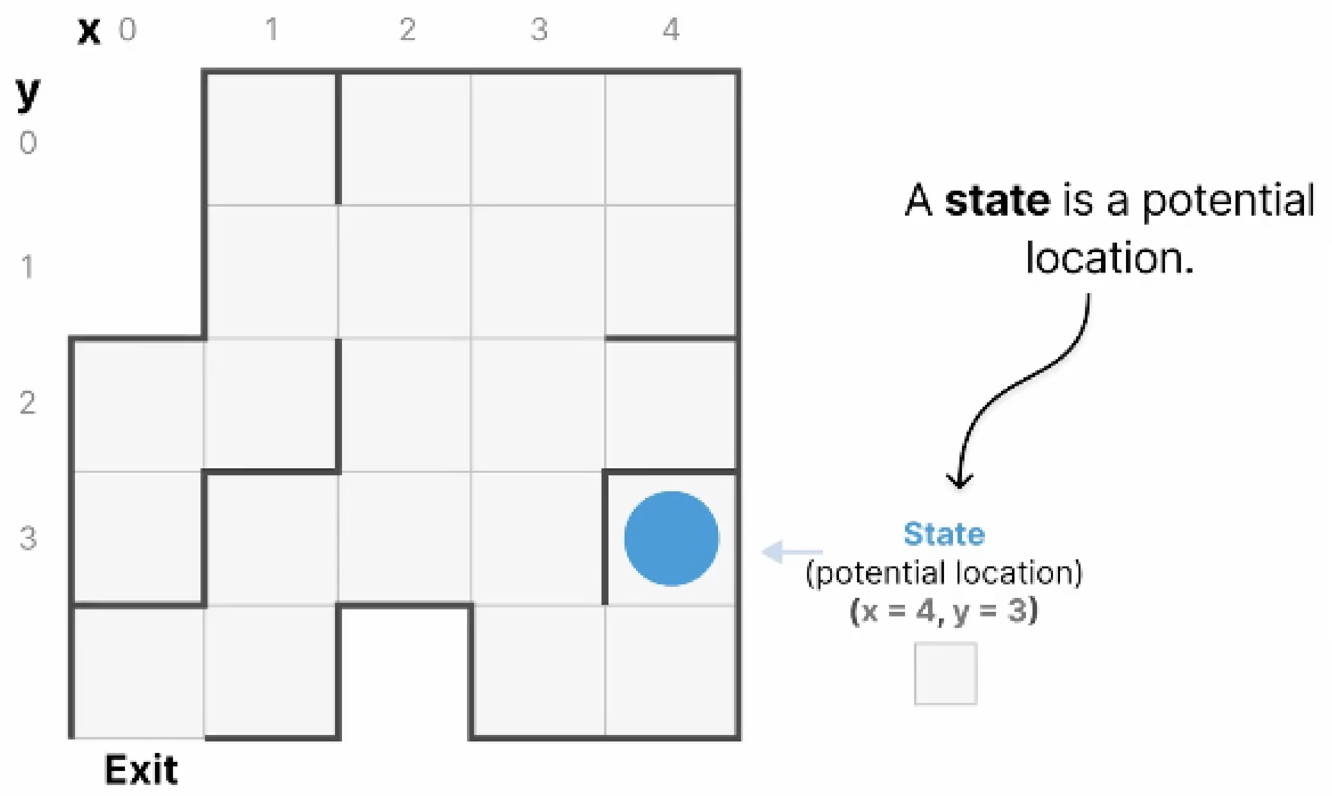



Transformer-based Models




State-Space Models

# State Space Model (SSM) Family

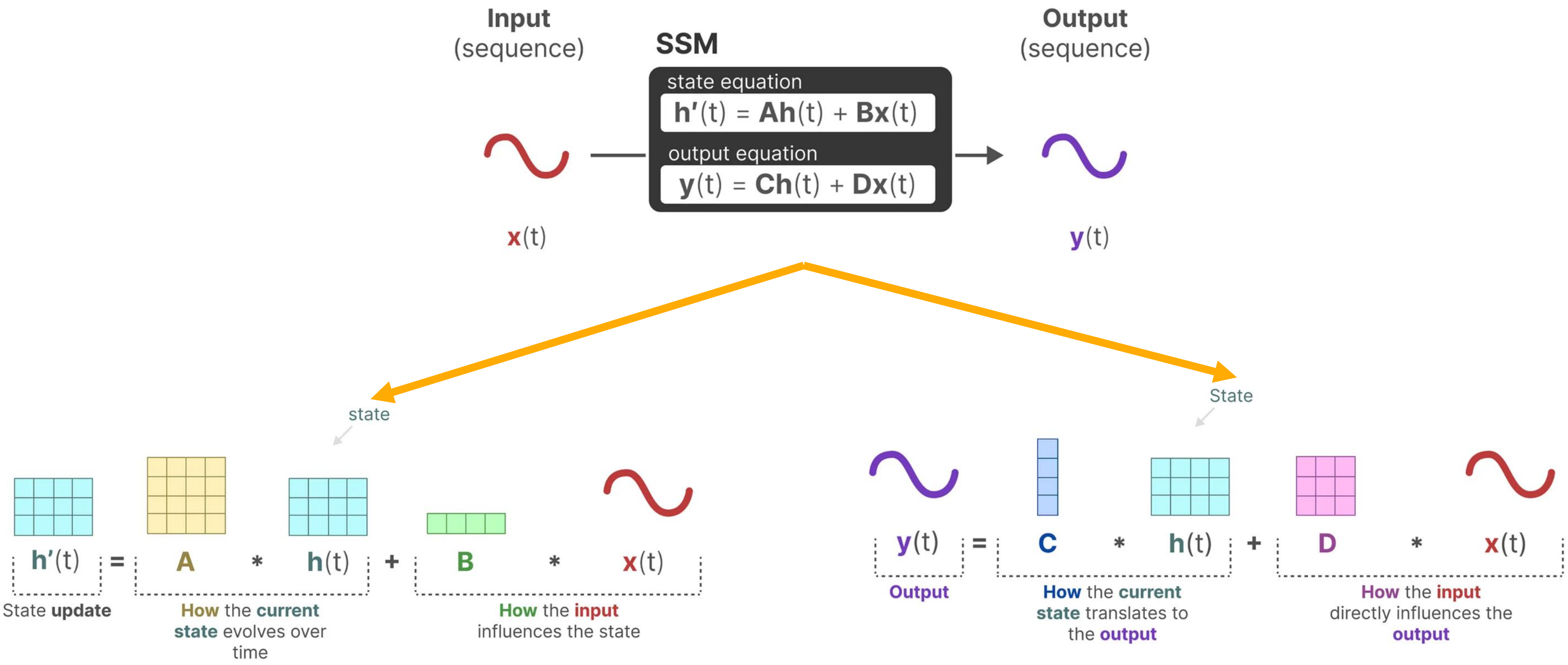




A State Space contains the minimum number of variables that fully describe a system. It is a way to mathematically represent a problem by defining a system's possible states.

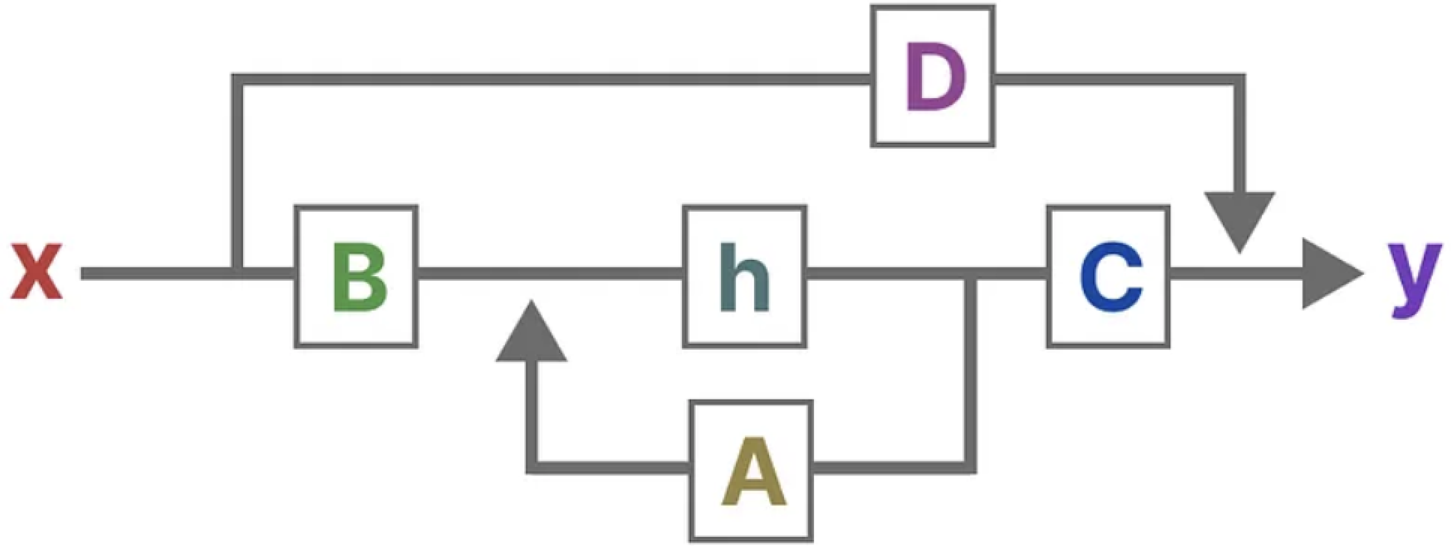
A Visual Guide to Mamba: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

# The Two Core Equations of SSMs



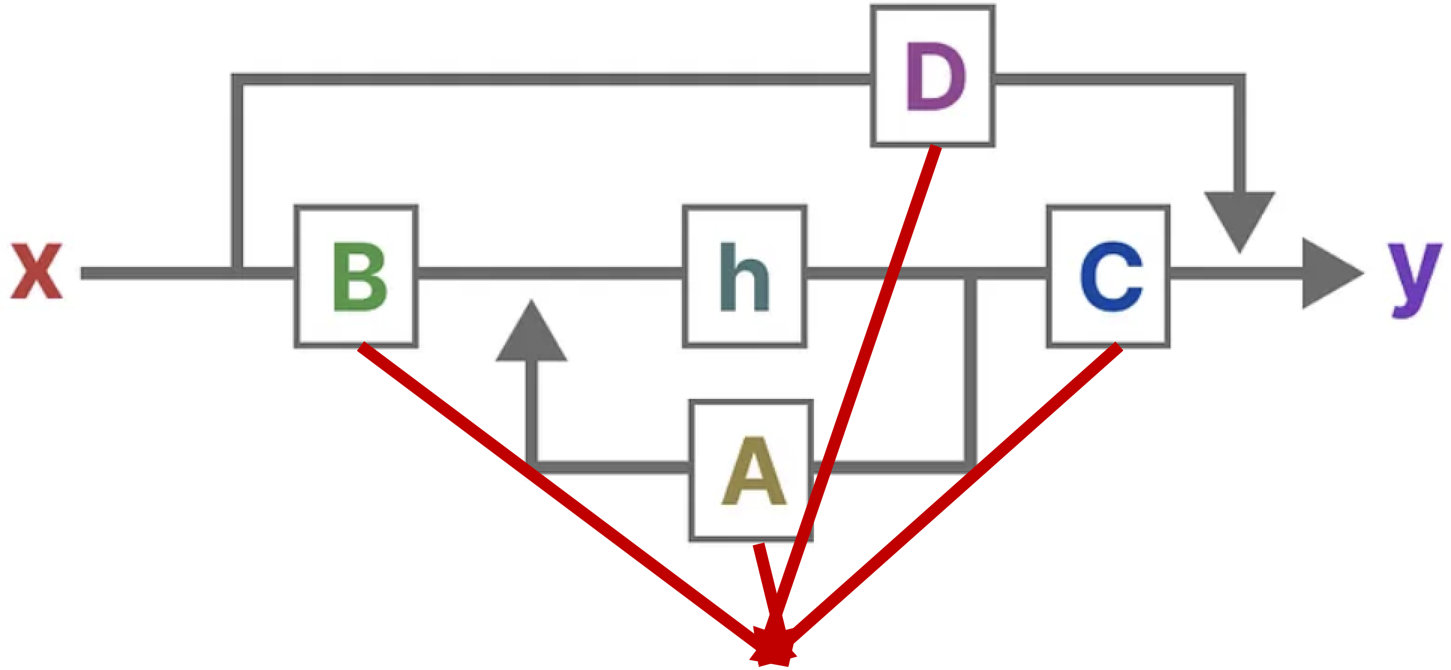

A Visual Guide to Mamba: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

# Simplified SSM Model Representation



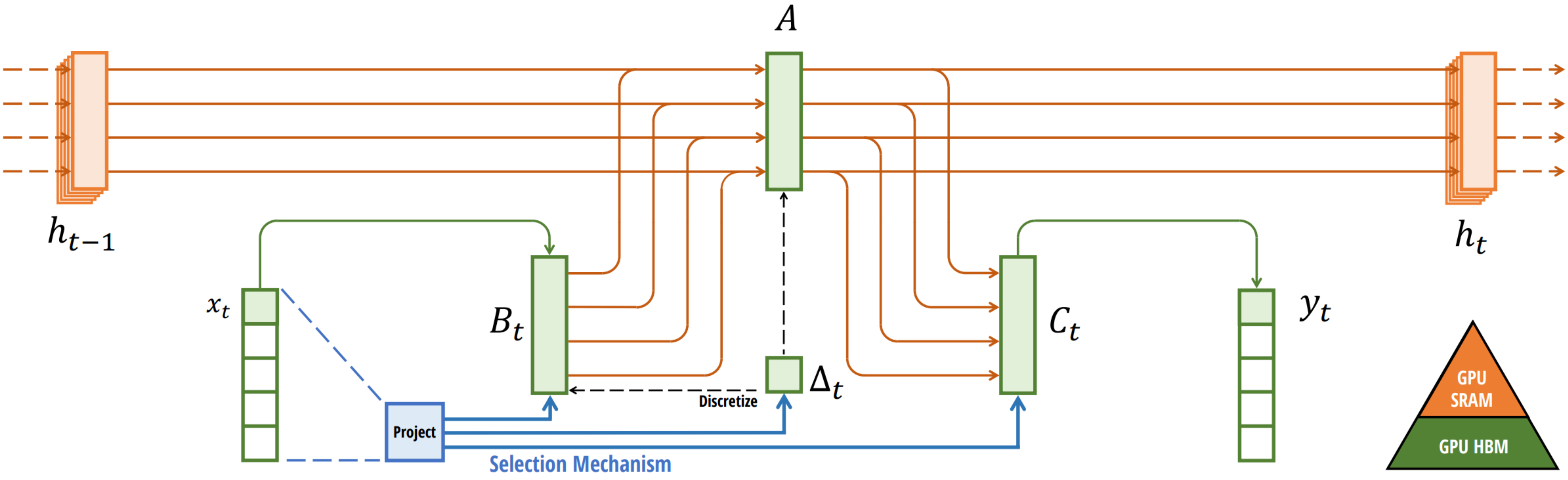

Continuous domain representation of SSMs

# Simplified SSM Model Representation




Constant with different inputs **X**

# Selective SSM or S6




Convert to discrete domain for LLMs!

A Visual Guide to Mamba: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

# Selective SSM or S6

## State space equation

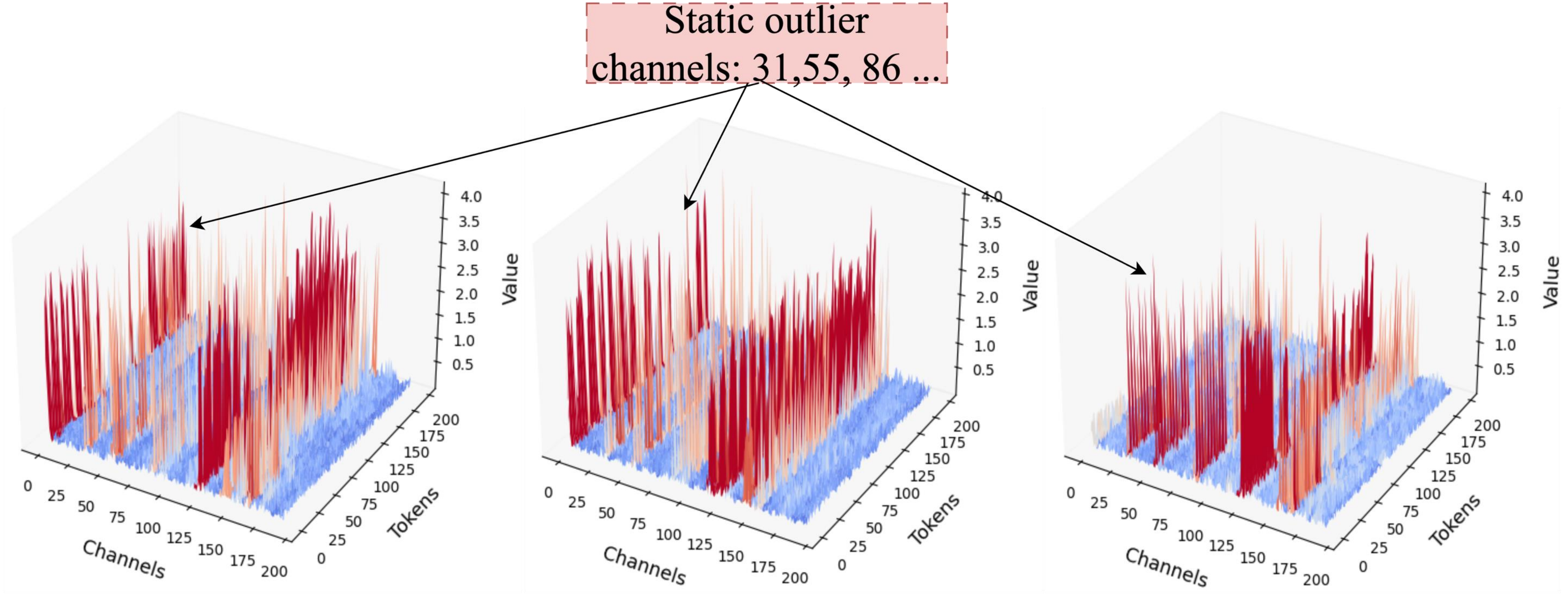
$$h(t) = \bar{A}(t) \odot h(t-1) + \bar{B}(t) \odot u(t); \;\; o(t) = C(t)h(t)$$

## Discretization

$$\bar{A}(t) = e^{(A\Delta(t))}; B(t) = W_B(u(t)); \bar{B}(t) = B\Delta(t)$$

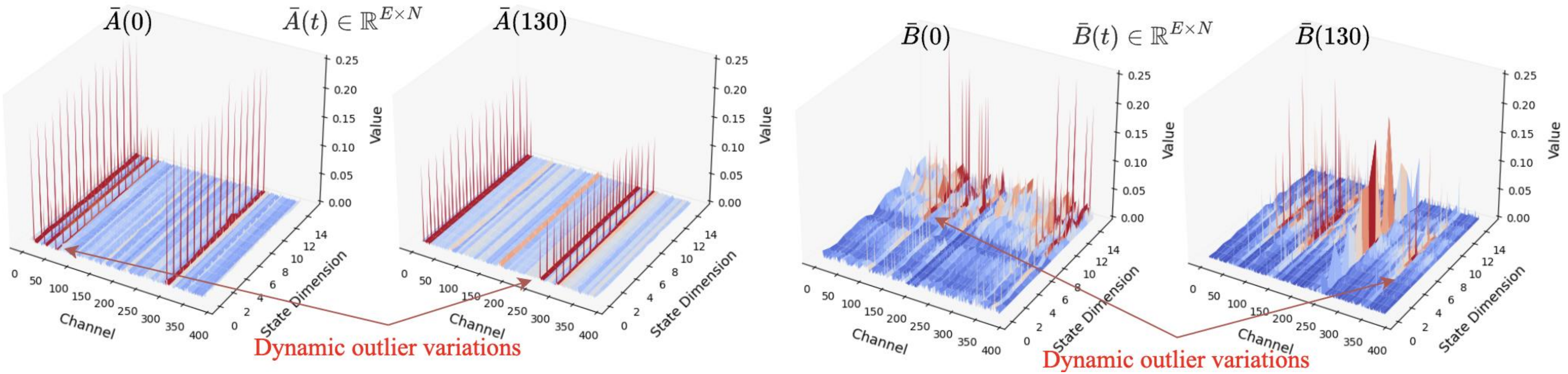
$$\Delta(t) = S^{+}(u(t)\Delta(t)_{\mathtt{proj}}); C(t) = (W_C(z(t)))^T$$

# Transformers' Static Outlier Patterns




DeiT Layer 9 for different inputs

# Dynamic Activation Variations



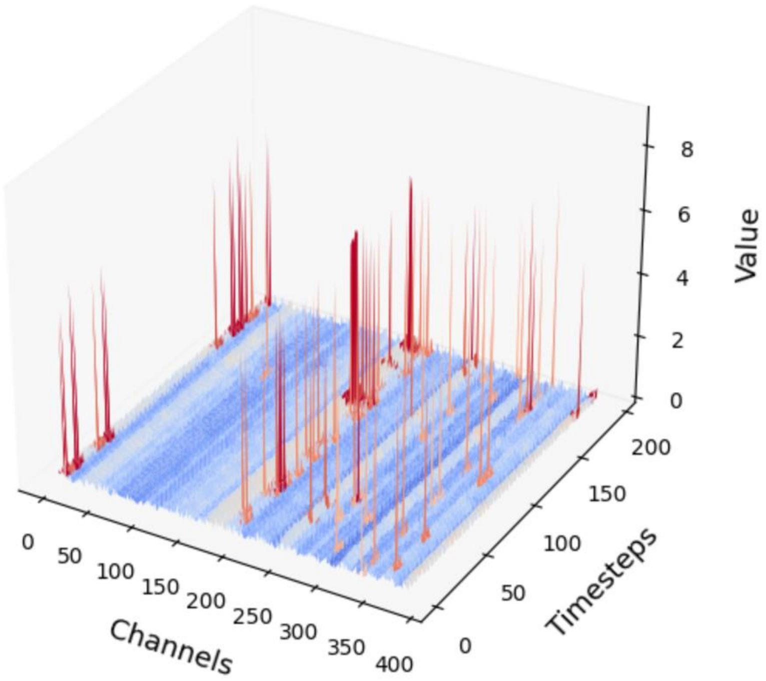

$$\bar{A}(t) = e^{(A\Delta(t))}; B(t) = W_B(u(t)); \bar{B}(t) = B\Delta(t)$$

$$\Delta(t) = S^{+}(u(t)\Delta(t)_{\texttt{proj}}); C(t) = (W_C(z(t)))^T$$

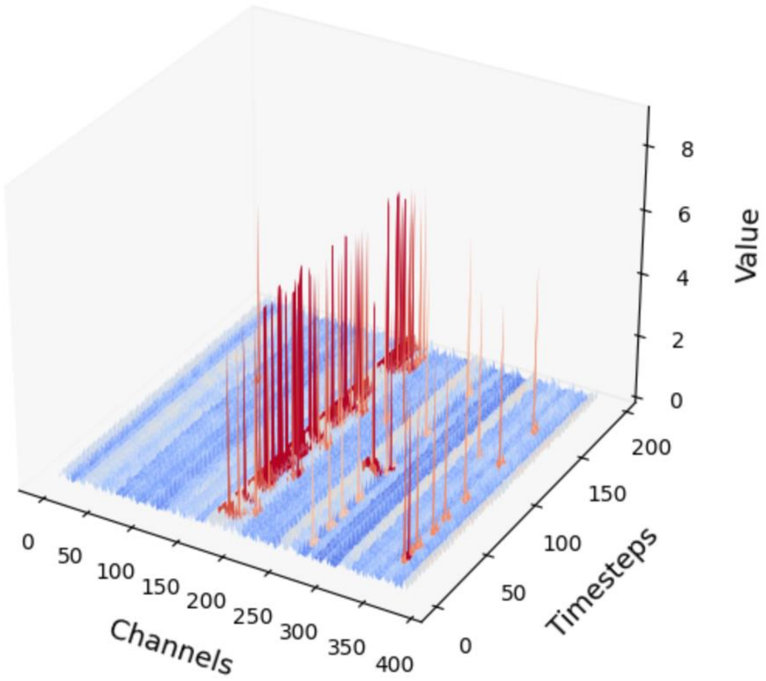
VMM activations exhibit dynamic inter-time-step channel variations

# Dynamic Activation Variations

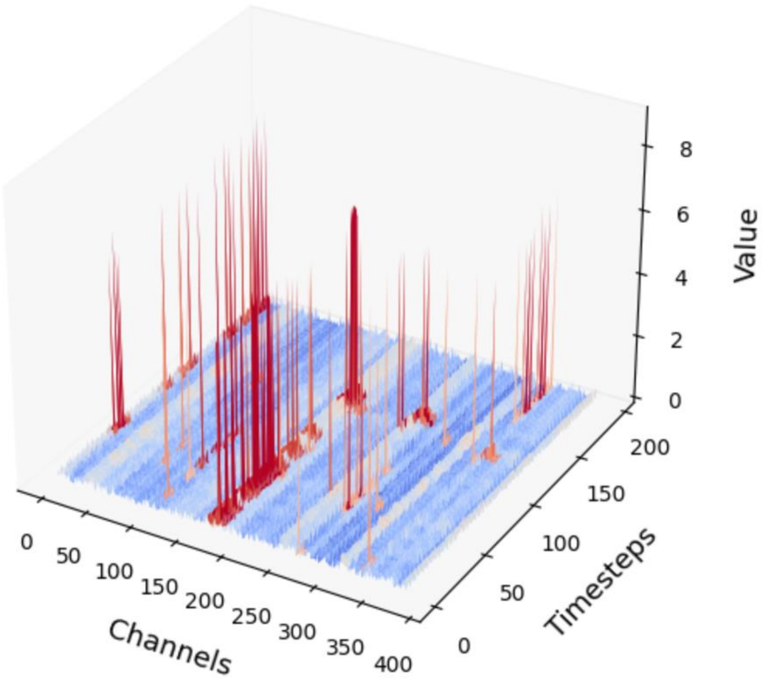
ViM-S Layer 3




$\Delta$




$A$



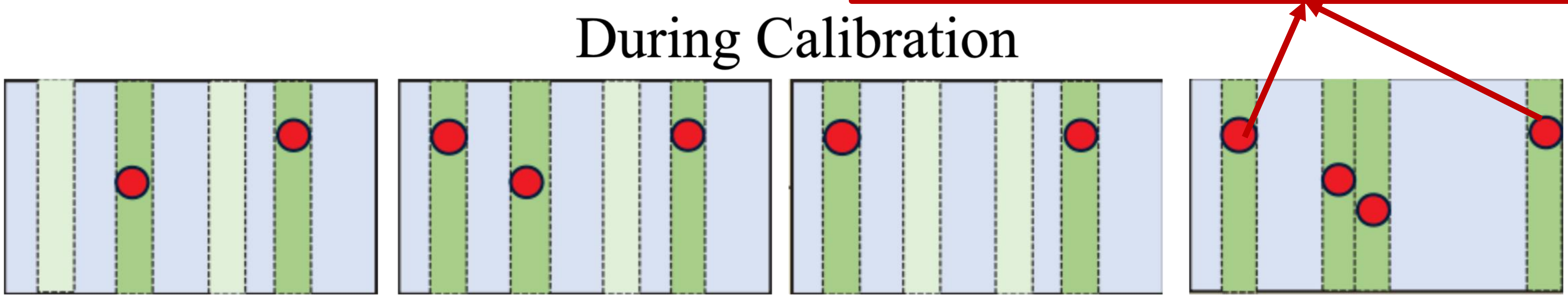

$B$

Dynamic variations across different activations too!

# Shortcoming of Static Determination

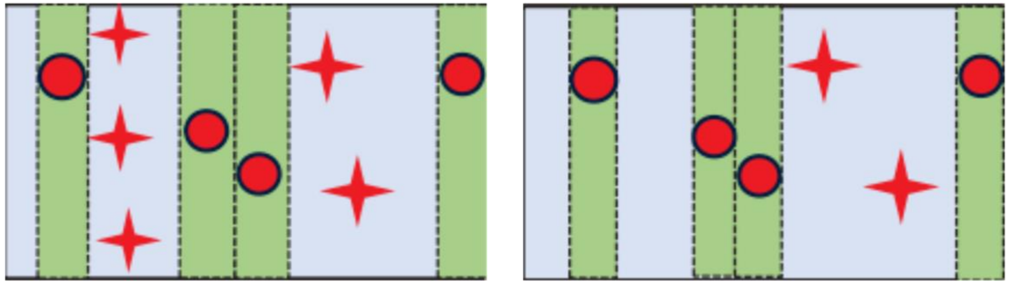
Overprovisioning for outliers during calibration

During Calibration



Time step 1 Time step 2 Time step 3 Time step 4

During inference



Time step 1 Time step 2

Misses new outlier channels

# OuroMamba-Quant

**Input :** Activation $X(t) \in \mathbb{R}^{N \times E}$, Static scale $S^I(t)$, Threshold $\theta$, Refresh rate $n_{\mathtt{refresh}}$, Outlier list $O_{\mathtt{list}}$, Inlier and outlier bit-precision $b_a^I, b_a^O$

**Output:** Quantized activation $X_q(t)$, Updated outlier list $O_{\mathtt{list}}$

**if** $t \; \% \; n_{refresh} == 0$ **then**

    $O_{\mathtt{list}} = \{\phi\}$

**end**

$S^D(t) = \mathtt{ComputeScale}(X(t)[:, c] \; \forall \, c \notin O_{\text{list}})$

**if** $S^D(t) > S^I(t)$ **then**

    **for** *each channel* $c$ *in* $X(t)$ ***not in*** $O_{list}$ **do**

        **if** $max(|X(t)[:, c]|) \geq \theta$ **then**

            $O_{\mathtt{list}} = O_{\mathtt{list}} \cup \{c\}$

        **end**

    **end**

**end**

$I(t), O(t) = \mathtt{Separate}(X(t), O_{\mathtt{list}})$

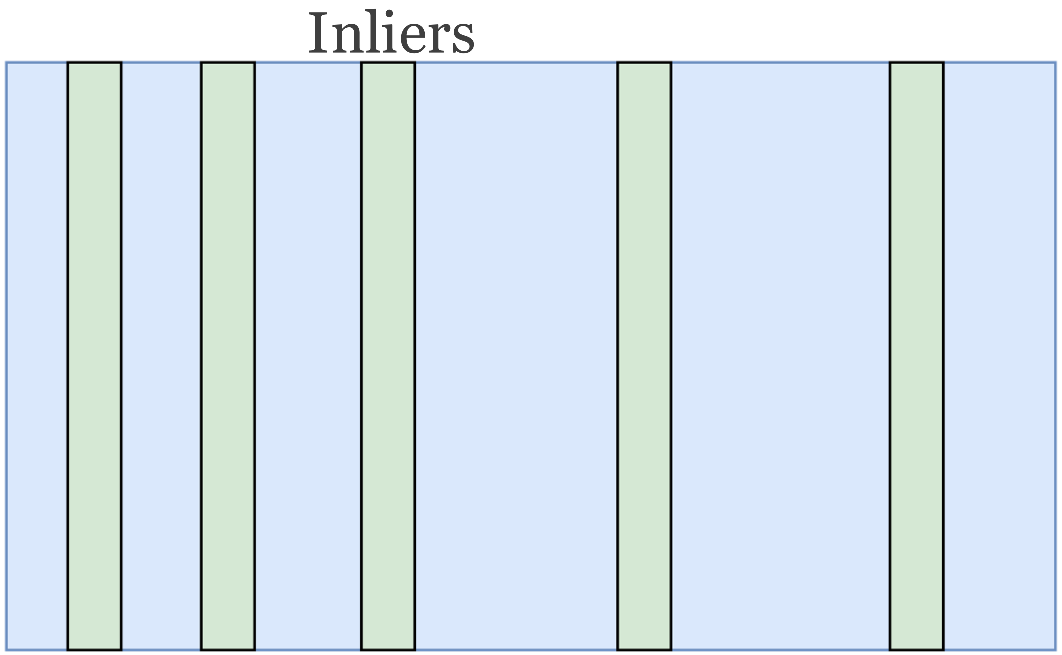
$I_q(t) = \mathtt{InlierQuant}(I(t), S^I(t), b_a^I)$

$O_q(t) = \mathtt{OutlierQuant}(O(t), b_a^O)$

$X_q(t) = \mathtt{Merge}(I_q(t), O_q(t))$

**return** $X_q(t), O_{\mathtt{list}}$

Outlier List: {}



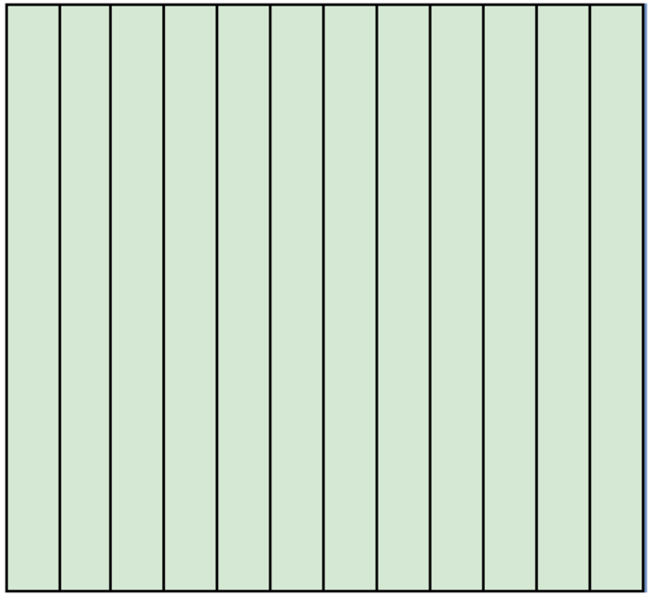

Offline: Determine threshold to detect outliers and inlier scale factor

# OuroMamba-Quant

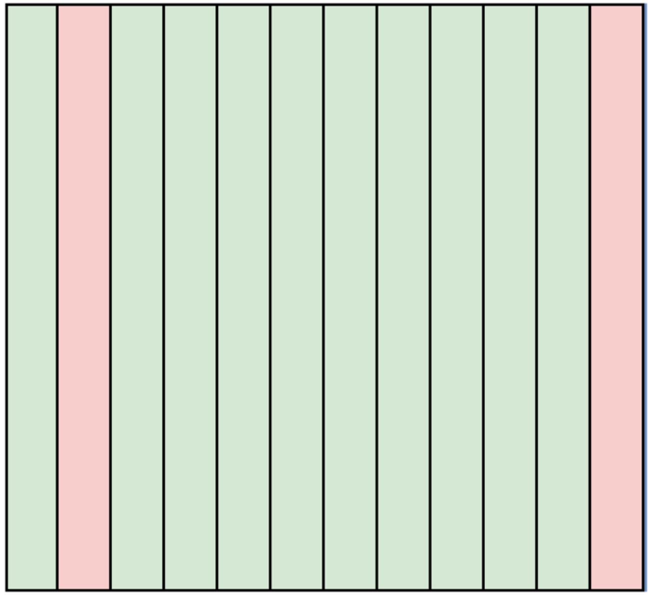
**Input :** Activation $X(t) \in \mathbb{R}^{N \times E}$, Static scale $S^I(t)$, Threshold $\theta$, Refresh rate $n_{\texttt{refresh}}$, Outlier list $O_{\texttt{list}}$, Inlier and outlier bit-precision $b_a^I, b_a^O$
**Output:** Quantized activation $X_q(t)$, Updated outlier list $O_{\texttt{list}}$

**if** $t \; \% \; n_{refresh} == 0$ **then**
    $O_{\texttt{list}} = \{\phi\}$
**end**
$S^D(t) = \texttt{ComputeScale}(X(t)[:, c] \; \forall \; c \notin O_{\text{list}})$
**if** $S^D(t) > S^I(t)$ **then**
    **for** *each channel $c$ in $X(t)$* ***not in*** $O_{list}$ **do**
        **if** $max(|X(t)[:, c]|) \geq \theta$ **then**
            $O_{\texttt{list}} = O_{\texttt{list}} \cup \{c\}$
        **end**
    **end**
**end**
$I(t), O(t) = \texttt{Separate}(X(t), O_{\texttt{list}})$
$I_q(t) = \texttt{InlierQuant}(I(t), S^I(t), b_a^I)$
$O_q(t) = \texttt{OutlierQuant}(O(t), b_a^O)$
$X_q(t) = \texttt{Merge}(I_q(t), O_q(t))$
**return** $X_q(t), O_{\texttt{list}}$

Outlier List: {}



Outlier List: {1,11}



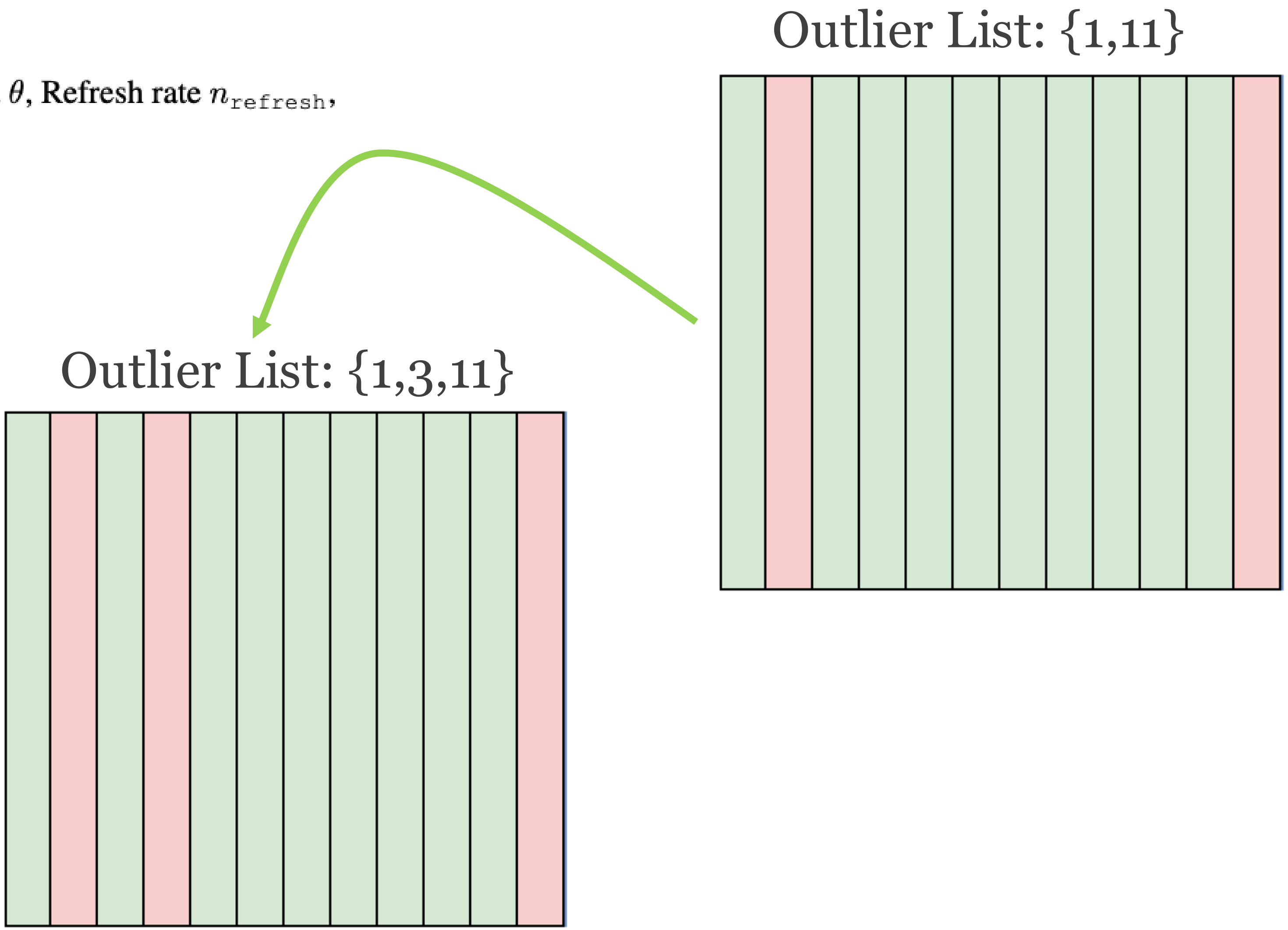
# OuroMamba-Quant

We must now identify which channel is the outlier channel

**Input :** Activation $X(t) \in \mathbb{R}^{N \times E}$, Static scale $S^I(t)$, Threshold $\theta$, Refresh rate $n_{\texttt{refresh}}$, Outlier list $O_{\texttt{list}}$, Inlier and outlier bit-precision $b_a^I, b_a^O$

**Output:** Quantized activation $X_q(t)$, Updated outlier list $O_{\texttt{list}}$

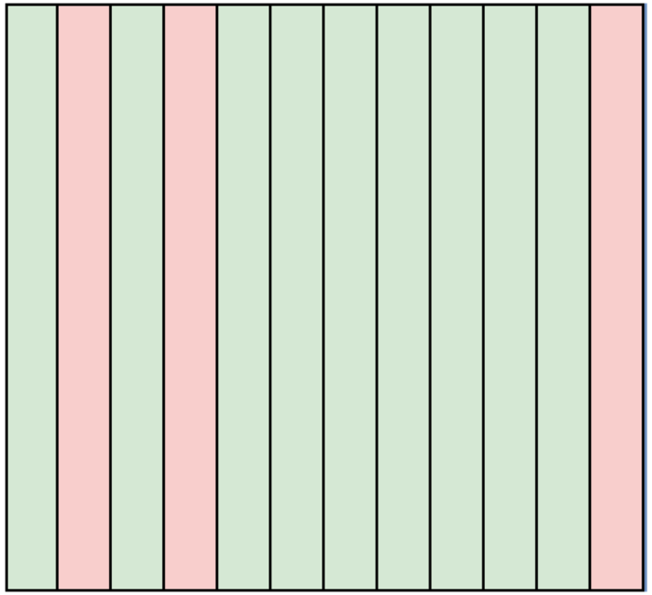
**if** $t \,\%\, n_{refresh} == 0$ **then**
    $O_{\texttt{list}} = \{\phi\}$
**end**
$S^D(t) = \texttt{ComputeScale}(X(t)[:, c] \;\forall\, c \notin O_{\text{list}})$
**if** $S^D(t) > S^I(t)$ **then**
    **for** *each channel c in* $X(t)$ ***not in*** $O_{list}$ **do**
        **if** $max(|X(t)[:, c]|) \geq \theta$ **then**
            $O_{\texttt{list}} = O_{\texttt{list}} \cup \{c\}$
        **end**
    **end**
**end**
$I(t), O(t) = \texttt{Separate}(X(t), O_{\texttt{list}})$
$I_q(t) = \texttt{InlierQuant}(I(t), S^I(t), b_a^I)$
$O_q(t) = \texttt{OutlierQuant}(O(t), b_a^O)$
$X_q(t) = \texttt{Merge}(I_q(t), O_q(t))$
**return** $X_q(t), O_{\texttt{list}}$



# OuroMamba-Quant

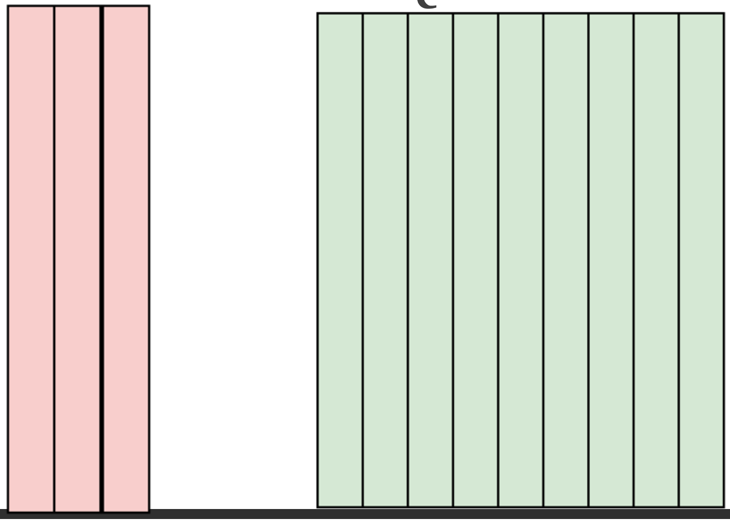
**Input :** Activation $X(t) \in \mathbb{R}^{N \times E}$, Static scale $S^I(t)$, Threshold $\theta$, Refresh rate $n_{\texttt{refresh}}$,
Outlier list $O_{\texttt{list}}$, Inlier and outlier bit-precision $b_a^I, b_a^O$
**Output:** Quantized activation $X_q(t)$, Updated outlier list $O_{\texttt{list}}$
**if** $t \ \% \ n_{refresh} == 0$ **then**
  $O_{\texttt{list}} = \{\phi\}$
**end**
$S^D(t) = \texttt{ComputeScale}(X(t)[:, c] \ \forall \ c \notin O_{\text{list}})$
**if** $S^D(t) > S^I(t)$ **then**
  **for** *each channel* $c$ *in* $X(t)$ ***not in*** $O_{list}$ **do**
    **if** $max(|X(t)[:, c]|) \geq \theta$ **then**
      $O_{\texttt{list}} = O_{\texttt{list}} \cup \{c\}$
    **end**
  **end**
**end**
$I(t), O(t) = \texttt{Separate}(X(t), O_{\texttt{list}})$
$I_q(t) = \texttt{InlierQuant}(I(t), S^I(t), b_a^I)$
$O_q(t) = \texttt{OutlierQuant}(O(t), b_a^O)$
$X_q(t) = \texttt{Merge}(I_q(t), O_q(t))$
**return** $X_q(t), O_{\texttt{list}}$

Outlier List: {1,3,11}



Outlier Quantization

Inlier Quantization



# OuroMamba-Quant

**Input** : Activation $X(t) \in \mathbb{R}^{N \times E}$, Static scale $S^I(t)$, Threshold $\theta$, Refresh rate $n_{\texttt{refresh}}$, Outlier list $O_{\texttt{list}}$, Inlier and outlier bit-precision $b_a^I, b_a^O$

**Output:** Quantized activation $X_q(t)$, Updated outlier list $O_{\texttt{list}}$

**if** $t \,\%\, n_{refresh} == 0$ **then**

  $O_{\texttt{list}} = \{\phi\}$

**end**

$S^D(t) = \texttt{ComputeScale}(X(t)[:, c] \;\forall\, c \notin O_{\text{list}})$

**if** $S^D(t) > S^I(t)$ **then**

  **for** *each channel* $c$ *in* $X(t)$ ***not in*** $O_{list}$ **do**

    **if** $max(|X(t)[:, c]|) \geq \theta$ **then**

      $O_{\texttt{list}} = O_{\texttt{list}} \cup \{c\}$

    **end**

  **end**

**end**

$I(t), O(t) = \texttt{Separate}(X(t), O_{\texttt{list}})$

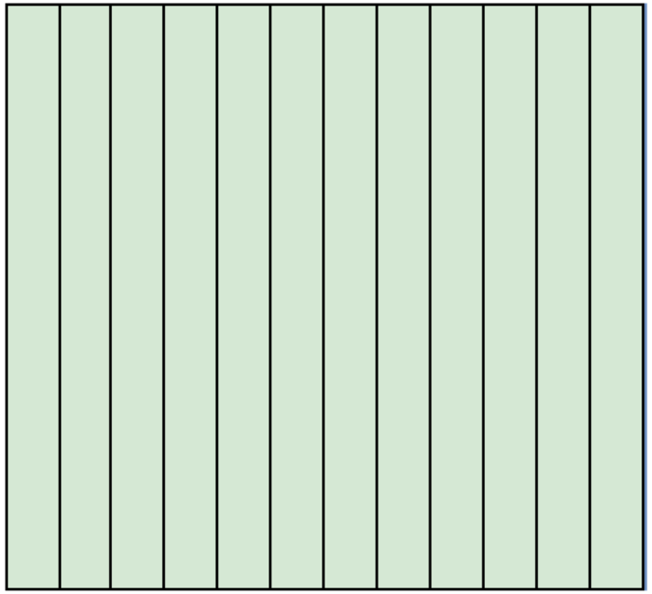
$I_q(t) = \texttt{InlierQuant}(I(t), S^I(t), b_a^I)$

$O_q(t) = \texttt{OutlierQuant}(O(t), b_a^O)$

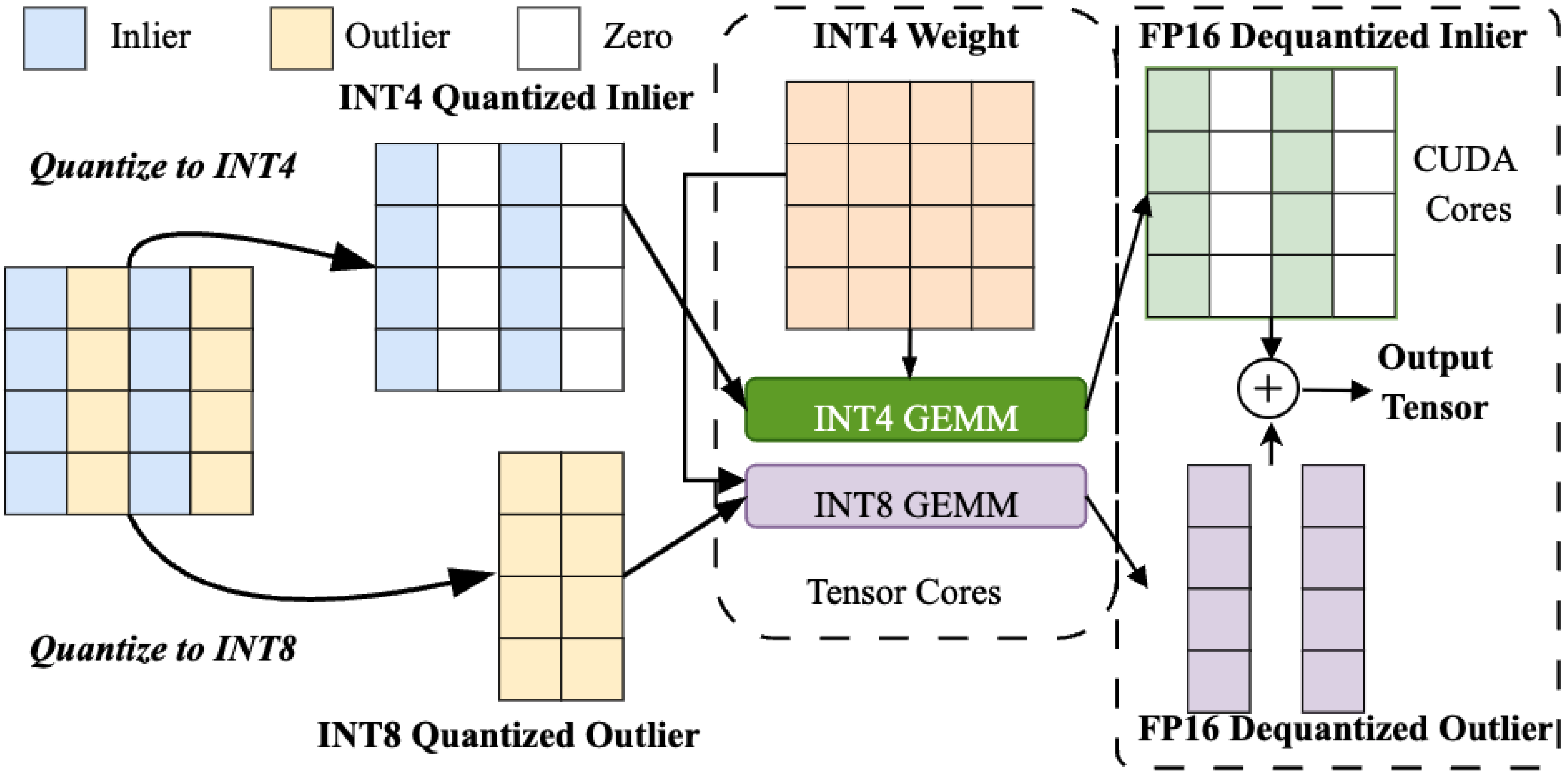
$X_q(t) = \texttt{Merge}(I_q(t), O_q(t))$

**return** $X_q(t), O_{\texttt{list}}$

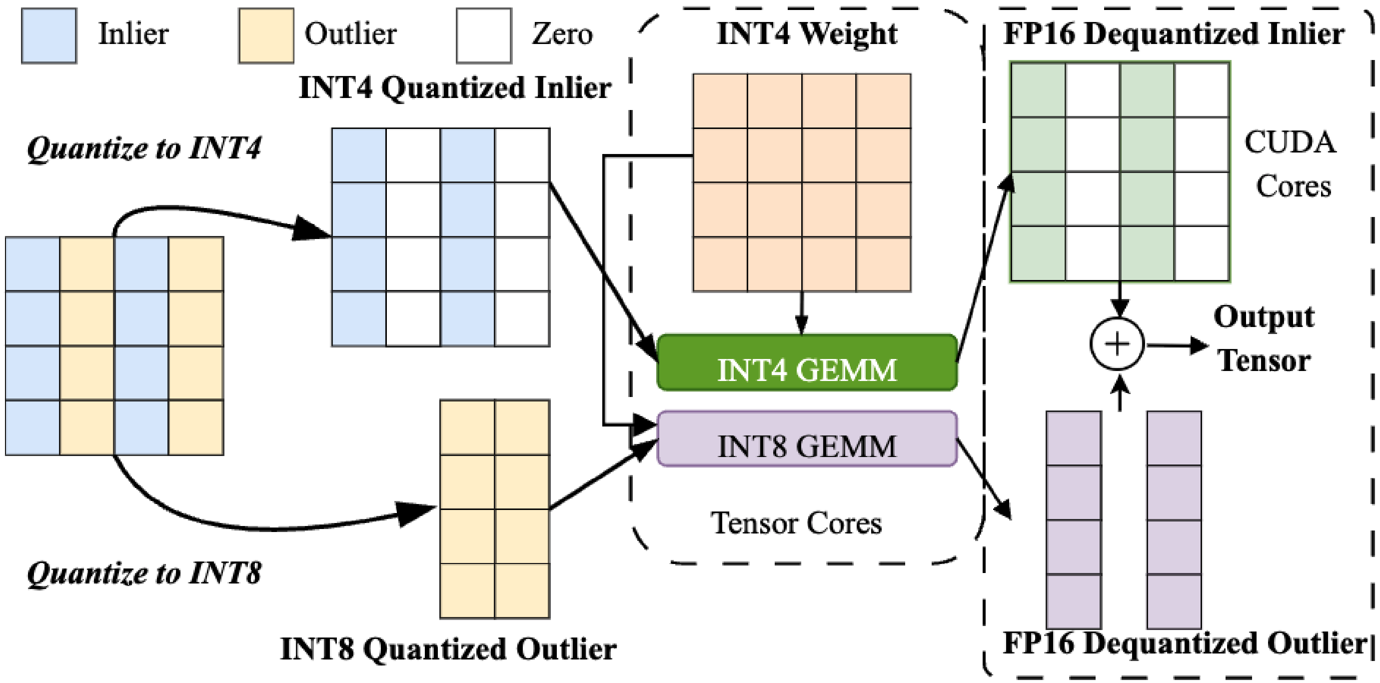
Outlier List: {}



# OuroMamba-Quant: W4A4 Hybrid GEMM

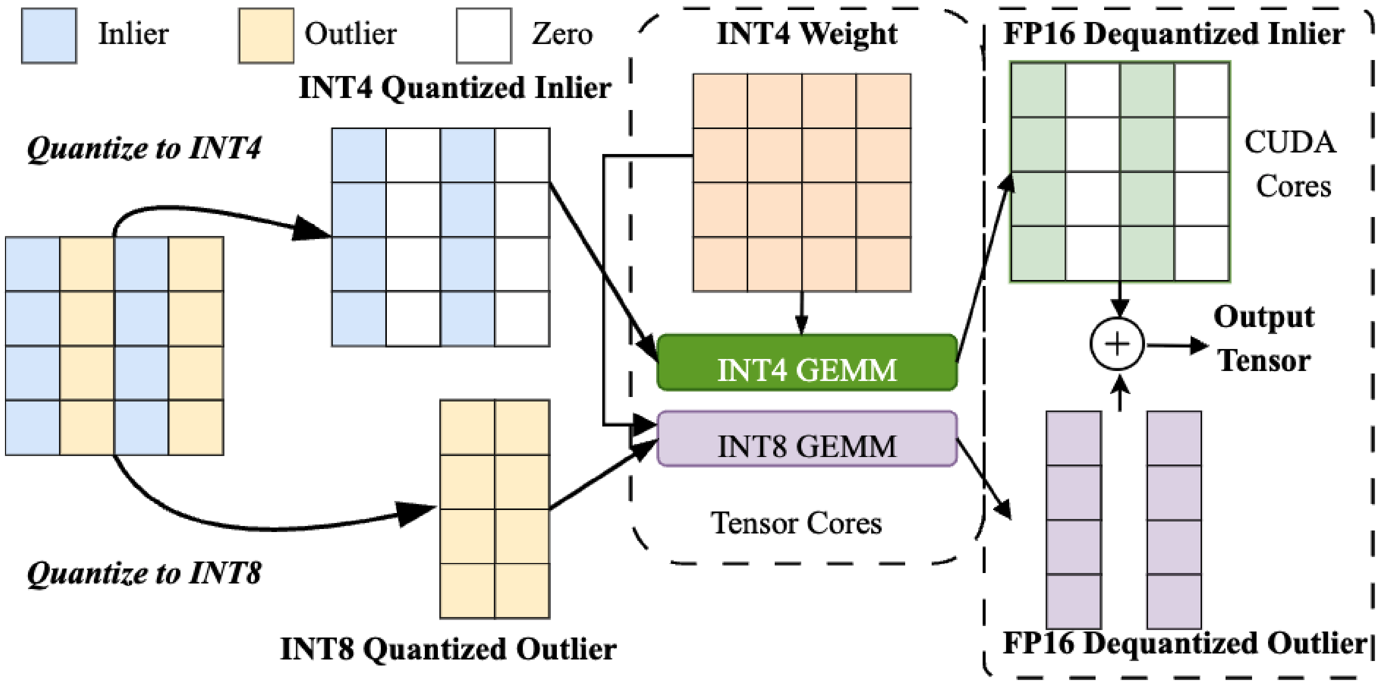


# OuroMamba-Quant: W4A4 Hybrid GEMM



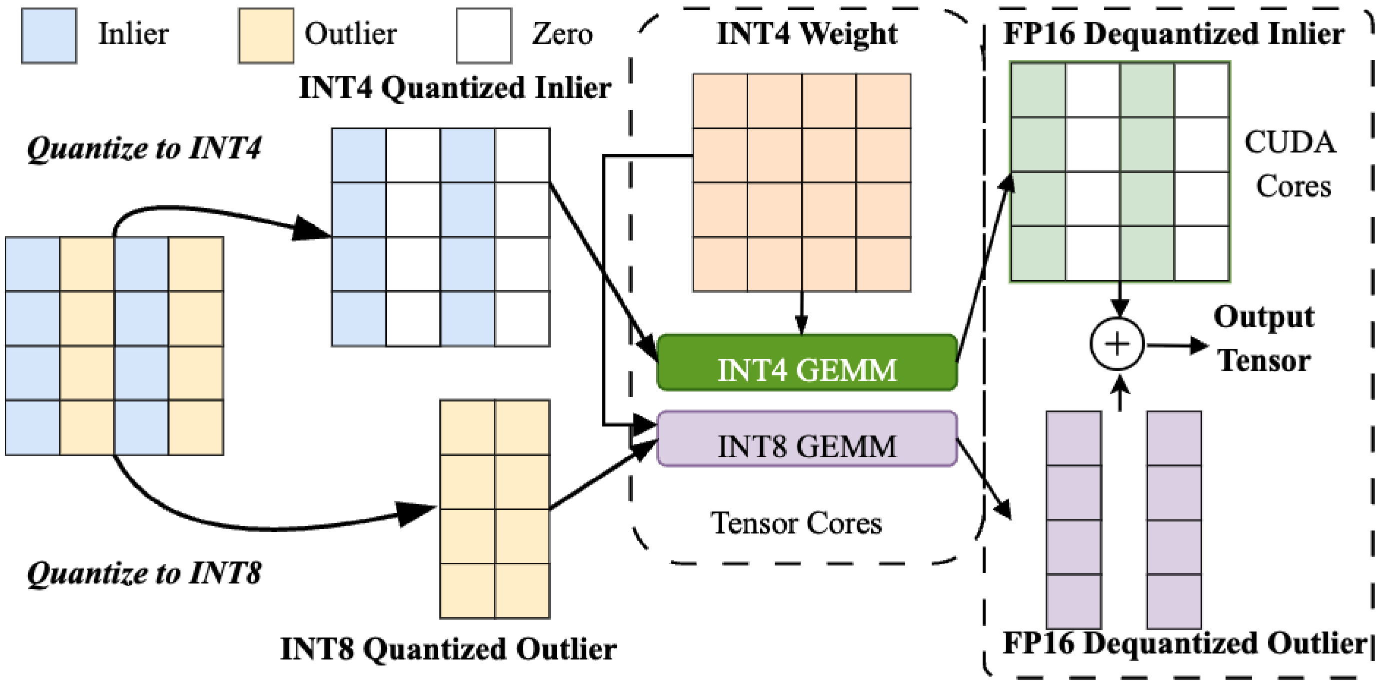

Activations are partitioned across channels and mapped to thread blocks, where each block independently compares its assigned channels against the threshold θ to identify outliers and updates the $O_{list}$

# OuroMamba-Quant: W4A4 Hybrid GEMM



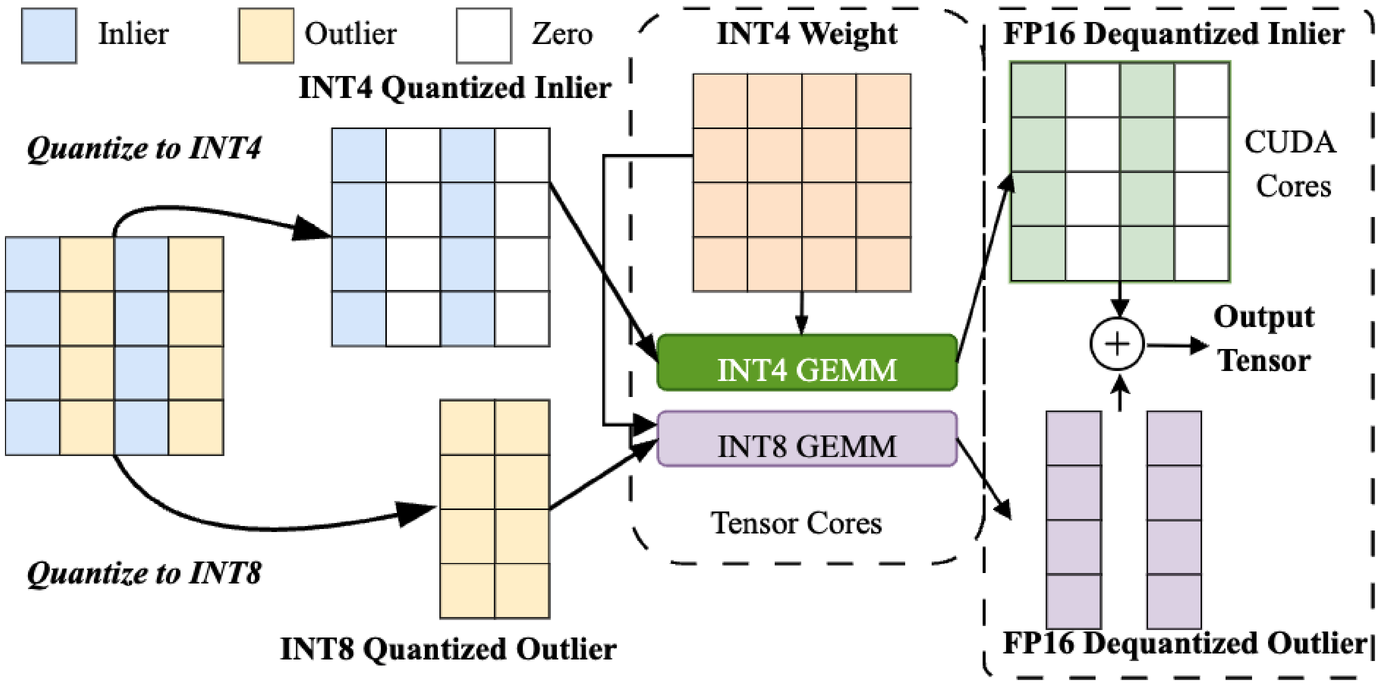

We pack two consecutive 4-bit inlier activations into one byte, with outlier positions set to zero, and leverage the **INT4 tensor cores for inlier GEMM**.

# OuroMamba-Quant: W4A4 Hybrid GEMM



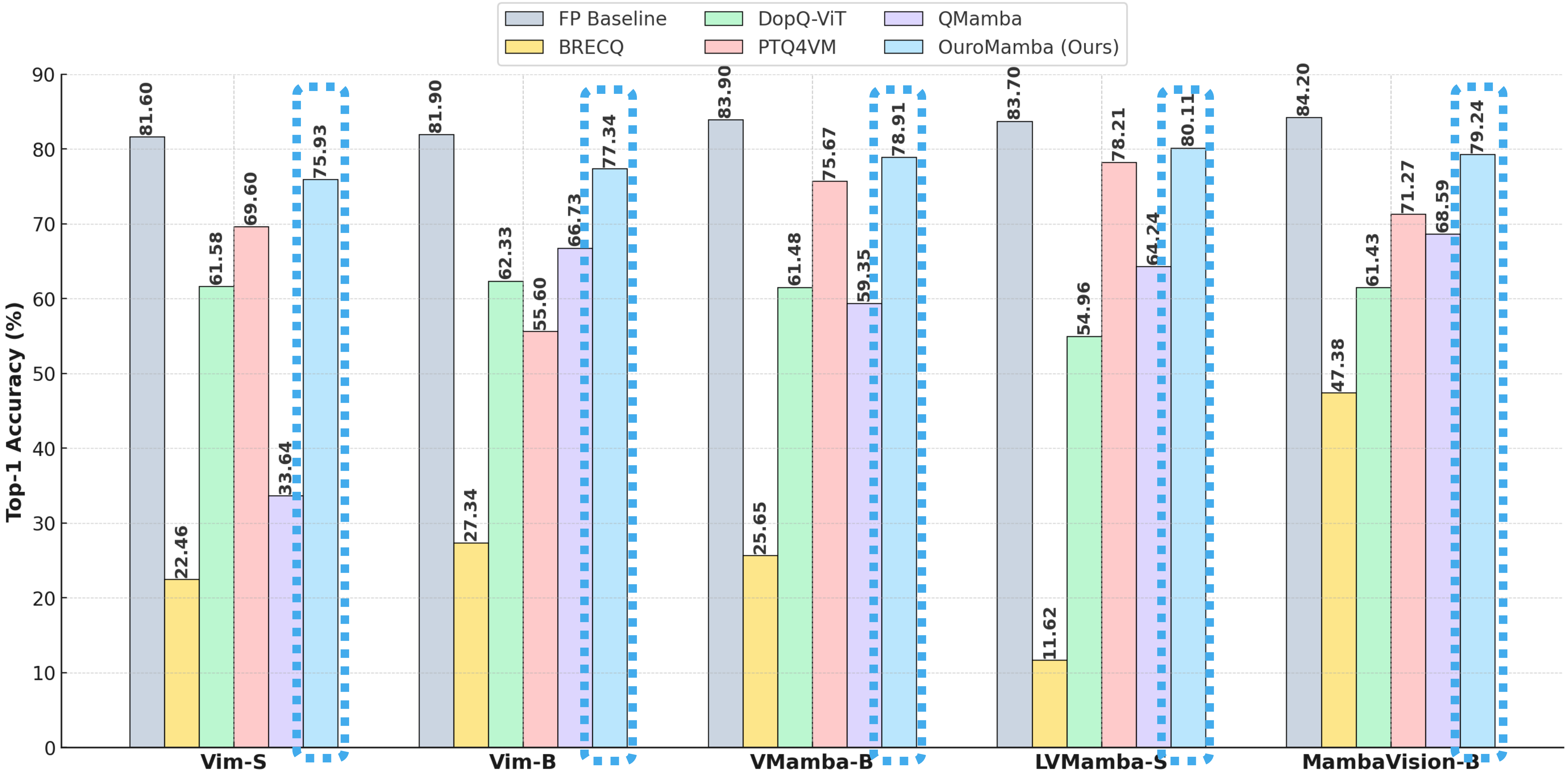

The outliers execute GEMM on **INT8 tensor cores**.

# OuroMamba-Quant: W4A4 Hybrid GEMM

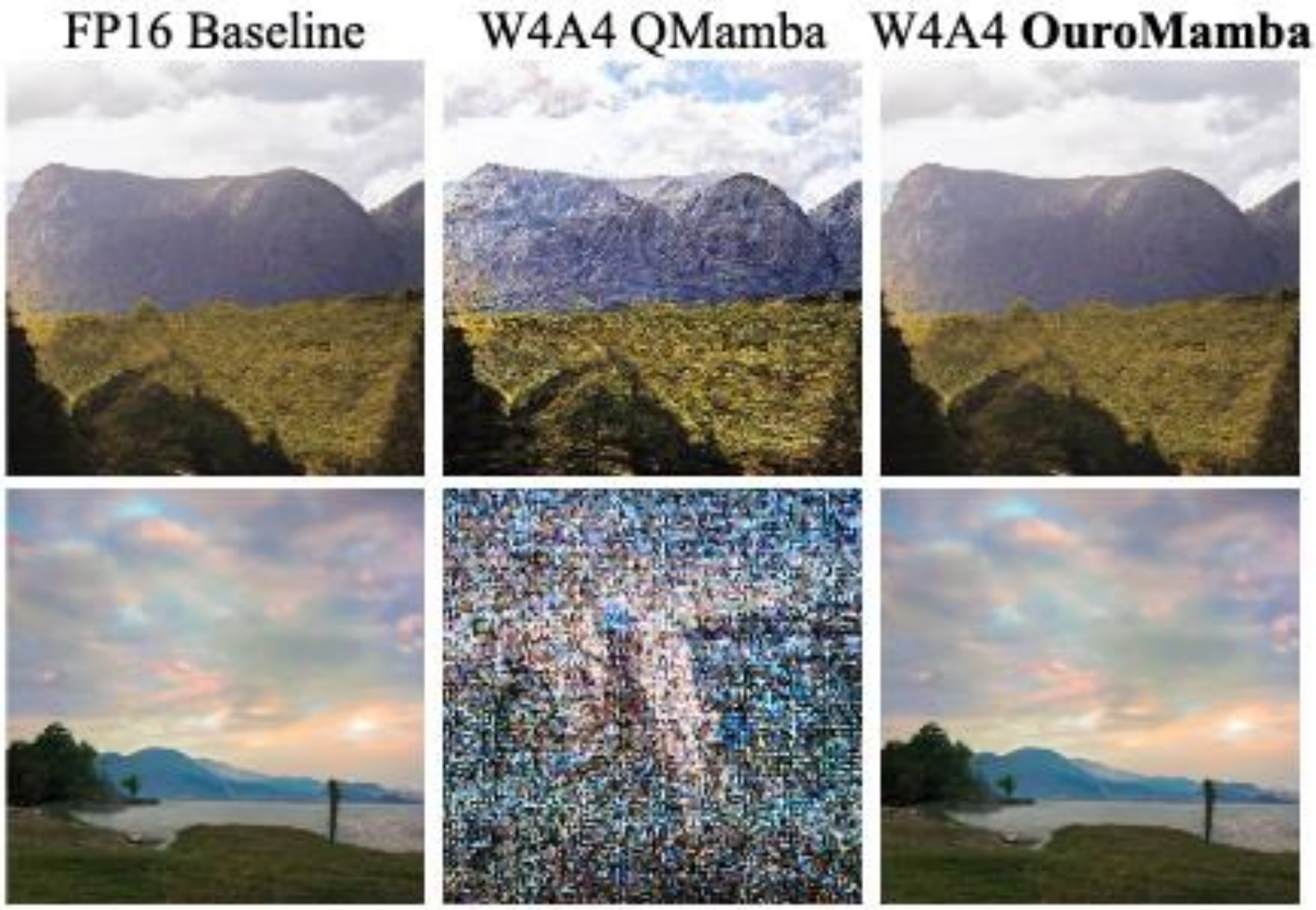



The outputs from the inlier and outlier GEMMs are dequantized to FP16 and summed together.

# Quantization Results: W4A4

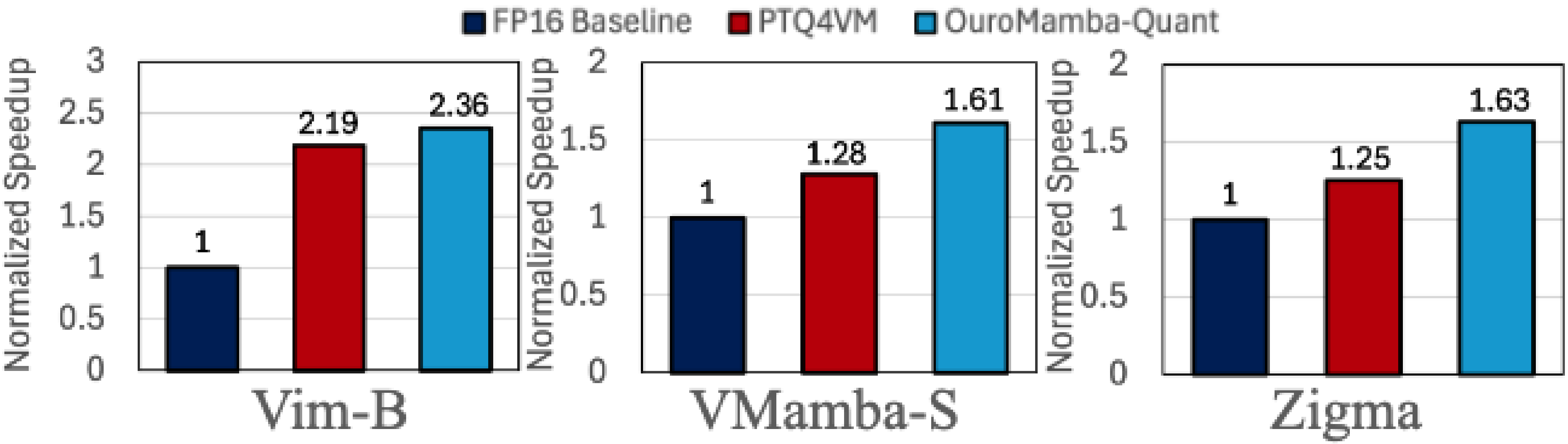


# Quantization Results for Diffusion Models




Zigma: SSM-based diffusion model

# End-to-End Latency Comparison



# Generalization to ViTs


Prompt: An astronaut relaxing on a beach chair, sipping coffee on Mars, with Earth visible in the sky

Text to image generation task