

## Problem

Time:

$t - 2$

$t - 1$

$t$

$t + 1$

Input:

The

cat

sat

?

## Persistent Memory

Llama

Mamba

$t - 2$

$[K_1, V_1]$

$h_1$

$t - 1$

$[K_1, V_1], [K_2, V_2]$

$h_2$

$t$

$[K_1, V_1], [K_2, V_2], [K_3, V_3]$

$h_3$

## Layer-wise Dataflow

Llama

$x(t)$

Q  
K  
V

KV  
Cache

SPDA

FFN

$y(t)$

Mamba

$x(t)$

Conv1D

$h(t)$

SSM

MLP

$y(t)$