FLAME: Fast Long-context Adaptive Memory for Event-based Vision

Biswadeep Chakraborty, Saibal Mukhopadhyay
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

The Problem: Event based Processing of Asynchronous Sparse Data

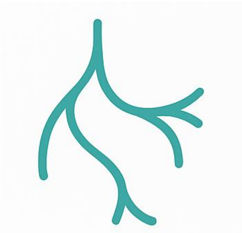
Low Latency Requirement
Real-time, Event by-Event processing

Context Retention
Adaptive Long-Range Memory

Computation Cost
Low Computation Cost

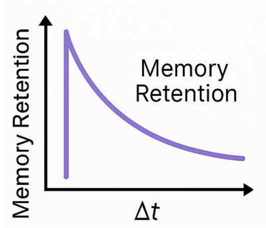
Novelty of FLAME

Multi - Timescale EAL



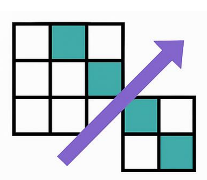
Extracts temporal features at multiple timescales

Event - Aware HiPPO



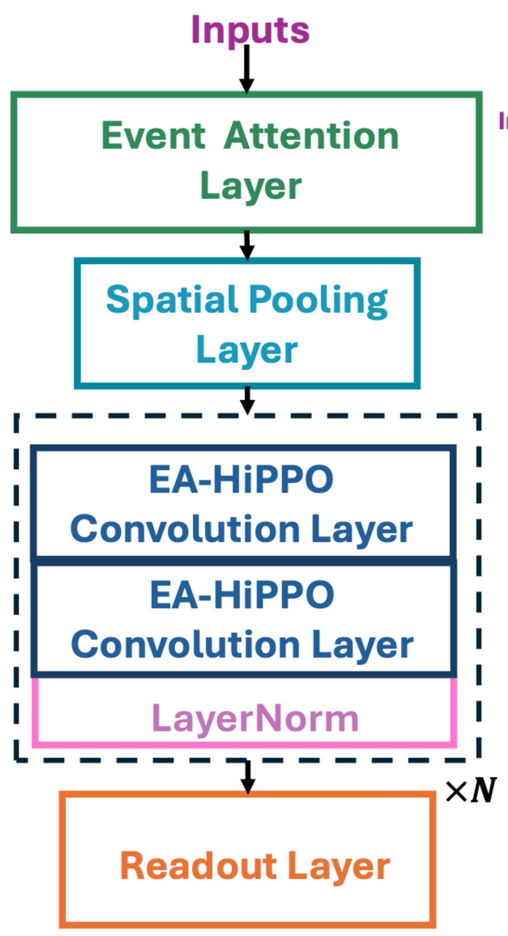
Adapts memory retention based on event sparsity

Efficient SSM

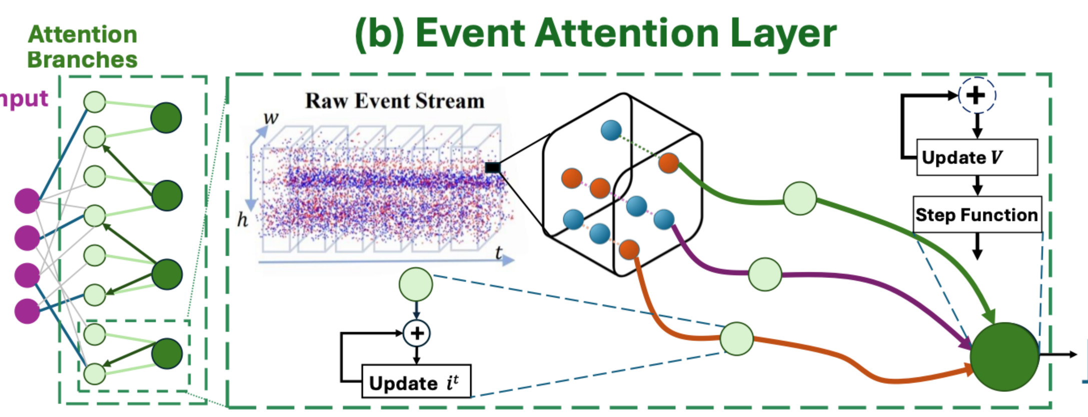


Uses NPLR + FFT for fast state-space updates

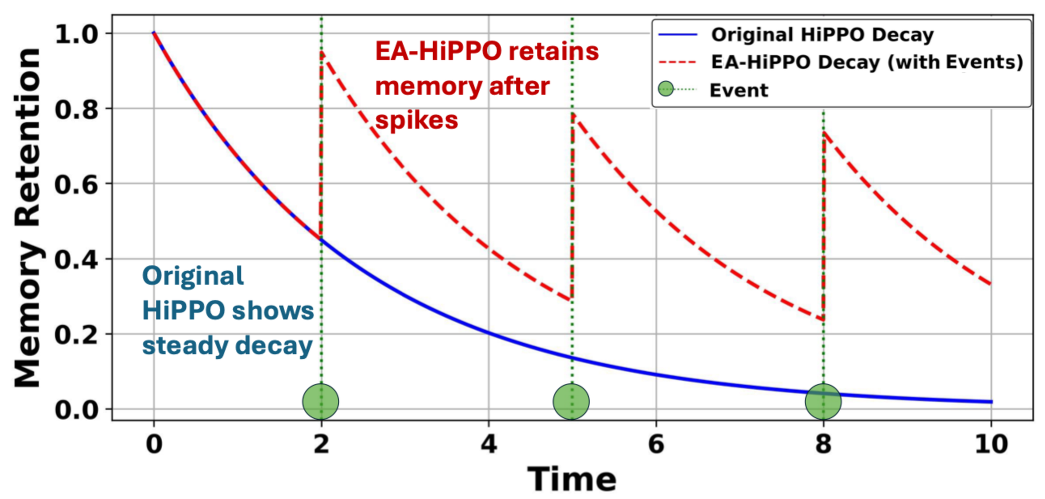
Block diagram of the proposed FLAME architecture



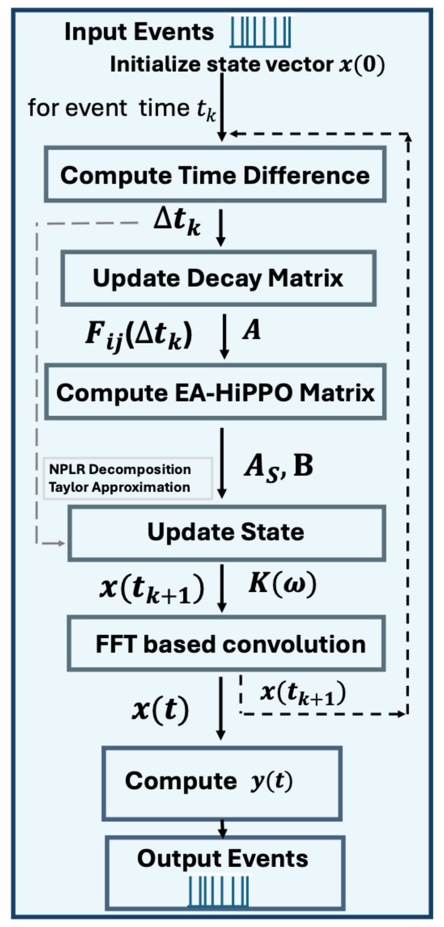
(a) Overall FLAMES Model



(b) Event Attention Layer



(c) Event Aware HiPPO



(d) EA-HiPPO Convolution Layer

(a) Overall architecture, combining neuro-inspired feature extraction with efficient state-space modeling.

(b) The Event Attention Layer (EAL) uses multi-branch Leaky Integrate-and-Fire (LIF) dynamics to extract multi-timescale temporal features from raw event streams.
(c) The core Event-Aware HiPPO (EA-HiPPO) dynamically modulates memory retention based on event timing (Δt), retaining context better than standard HiPPO after sparse events.

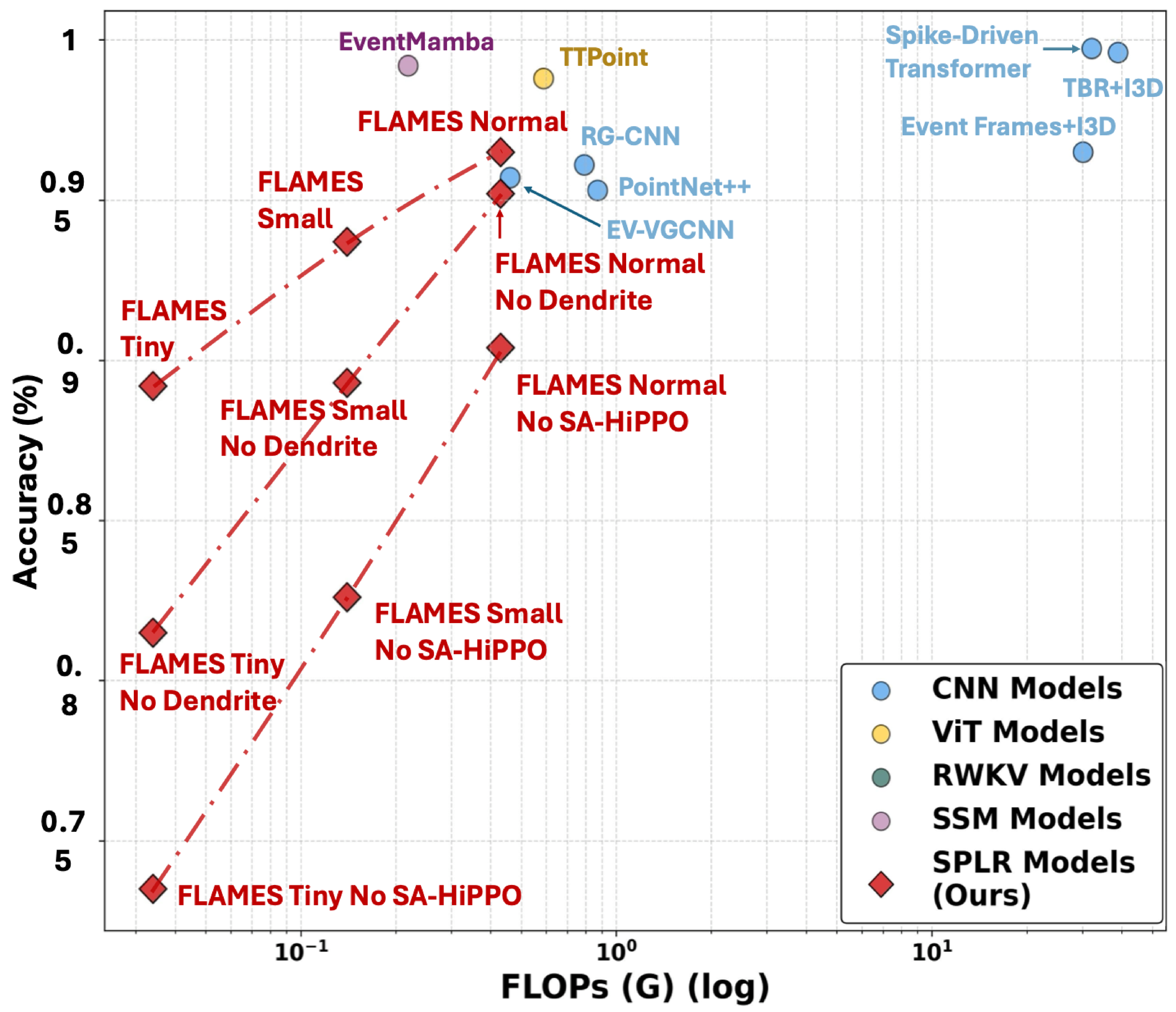
(d) The EA-HiPPO Convolution Layer achieves efficiency via asynchronous updates, Normal-Plus-Low-Rank (NPLR) decomposition, and FFT-based convolution.

Acknowledgement

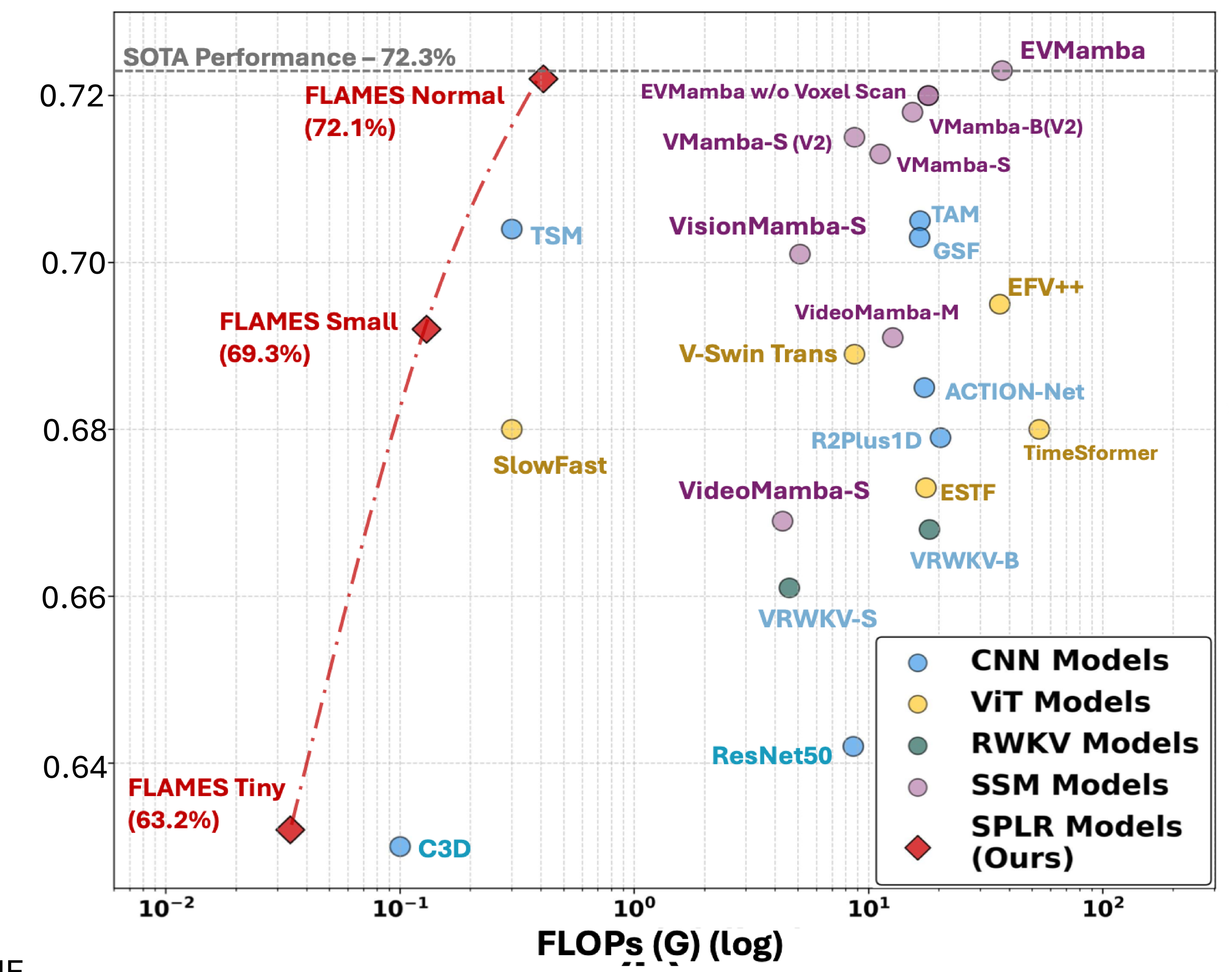
CogniSense Center for Cognitive Multispectral Sensors
LPS The Laboratory for Physical Sciences
Semiconductor Research Corporation

Comparing FLAME variants with other State-of-the-Art (SOTA) models

Accuracy versus GFLOPs across various event-based vision datasets



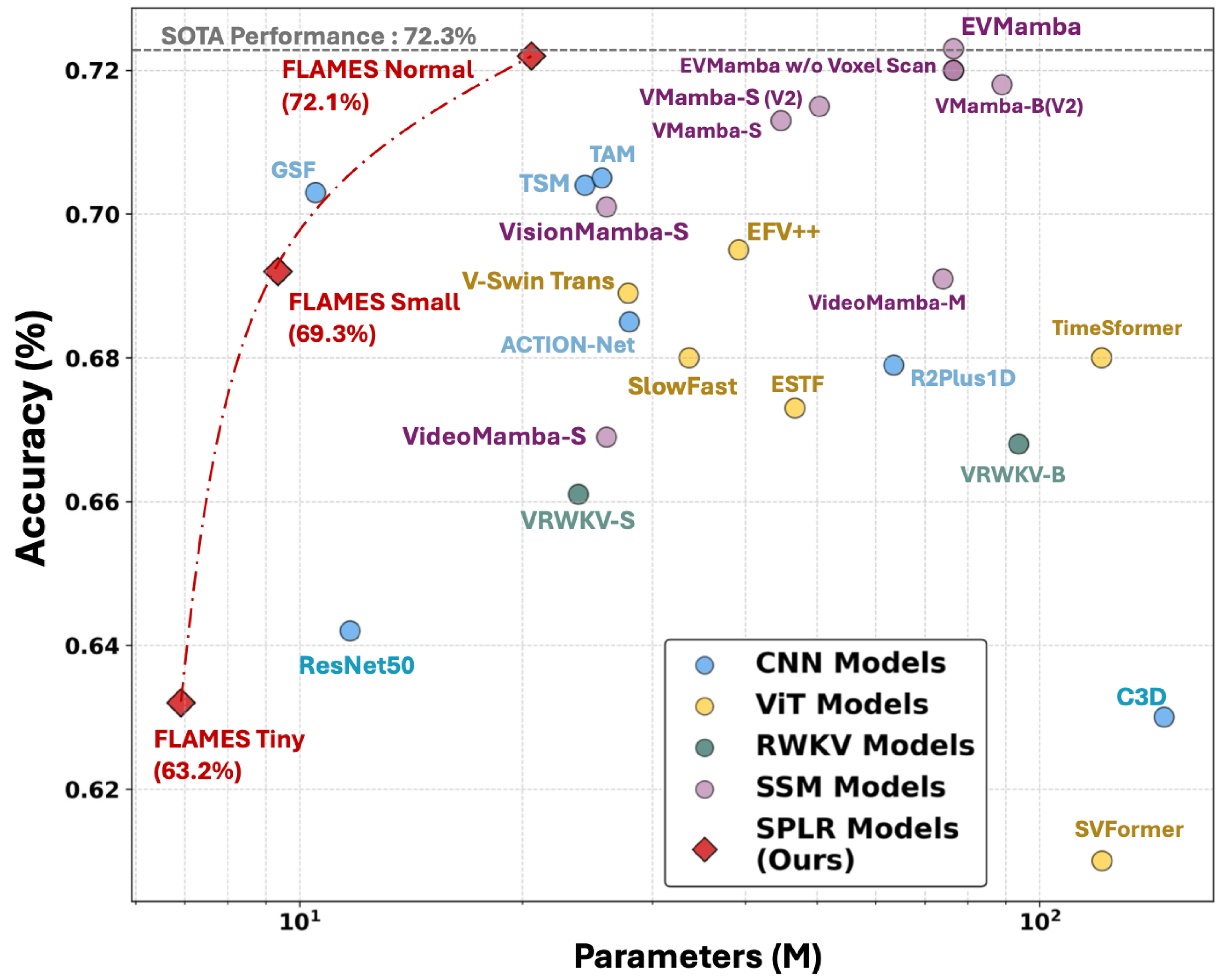
Performance on **DVSGesture128**, including ablation studies for FLAME demonstrating the impact of removing the Event Attention Layer (No Dendrite) or replacing EA-HiPPO with standard LIF neurons (No SA-HiPPO).



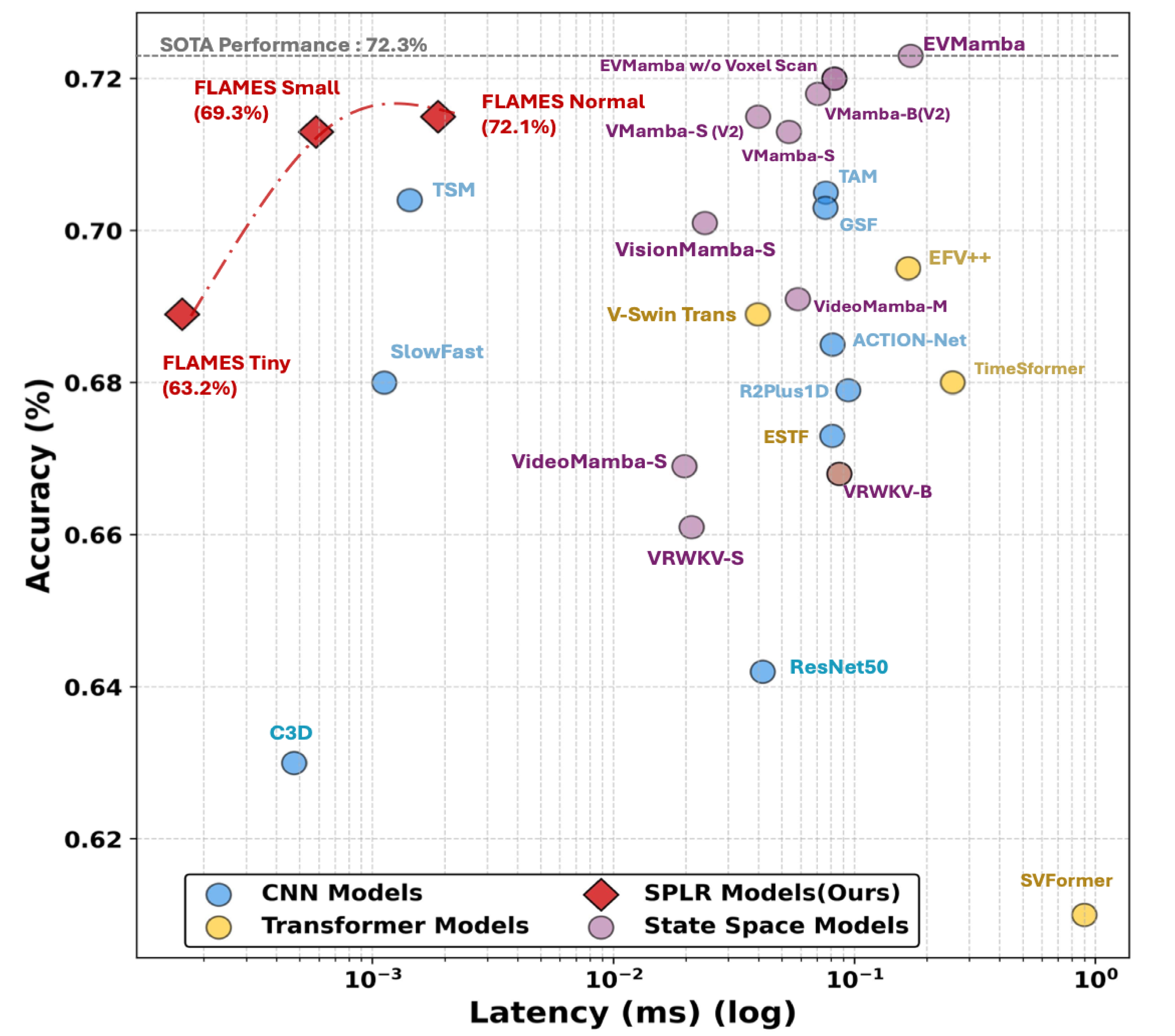
Performance on the high-resolution **CeleX-HAR** dataset, showcasing FLAME's efficiency at scale.

FLAME variants demonstrate a superior trade-off compared to SOTA models

Efficiency analysis on the CeleX-HAR dataset, measured on an NVIDIA A100 GPU



Accuracy versus Parameters (M): FLAME achieves competitive accuracy with significantly lower parameter counts than many high-performance models.



Accuracy versus Inference Latency (ms) (log scale): FLAME models exhibit substantially lower latency, confirming the efficiency of the asynchronous, event-by-event design for real-time applications.