A Neuromorphic Approach to Event-Based Vision: From Encoding to Adaptive Architectures

Priya Panda*, **Assistant Professor,**

Visiting Faculty@Google DeepMind

Students: Ginny Xiao, Donghyun Lee, Wei Fang, Yuhang Li

Electrical & Computer Engineering, Yale Unive

Email: priya.panda@yale.edu

*** Moving to University of Southern California (USC) as Associate Professor in Aug.'25**

Yale University

Intelligent Computing Lab

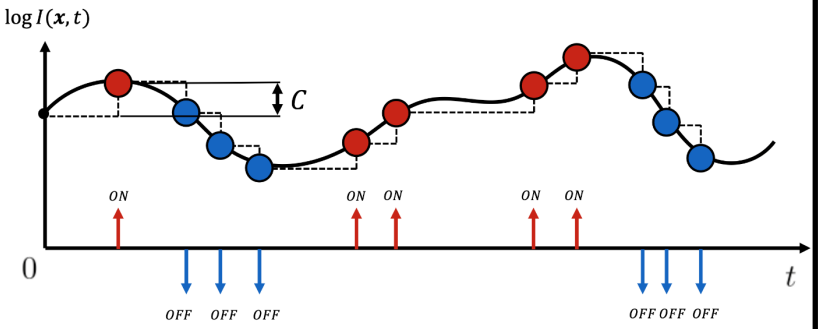
https://intelligentcomputinglab.yale.edu/

Event Data: Background

What is an Event Camera?

- Novel sensor that measures only **brightness changes (motion)**
- An event is triggered at a **pixel** if

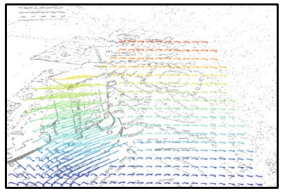
$$\log I(\boldsymbol{x}, t) - \log I(\boldsymbol{x}, t - \Delta t) = \pm C$$



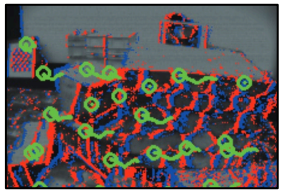
Applications

Vision

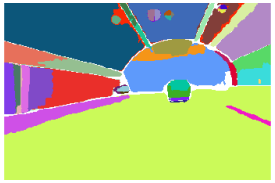
- Optical flow estimation [3]



- Object tracking [4]



- Semantic Segmentation [5]



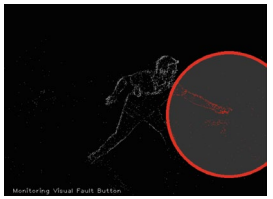
Autonomous Driving

- Keyword spotting [6]

- Driver assistant system [7]

Industry

- Detect dangerous scene [8]



Yale University

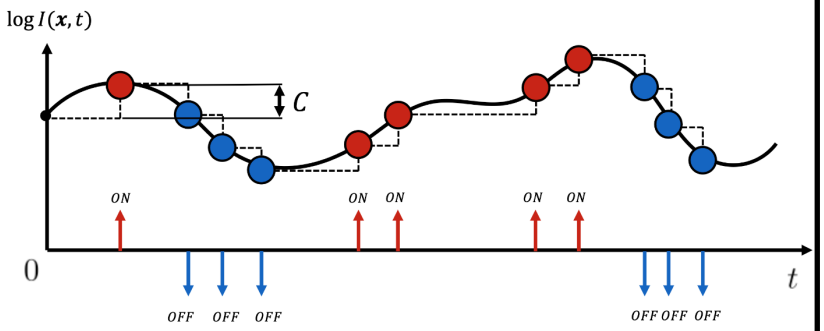
[1] Scaramuzza. Event_Cameras_Tutorial.pdf
[2] Video from https://youtu.be/LauQ6LWTkxM?t=30
[3] Hamann, Friedhelm, et al. ECCV, 2024
[4] Ikura, Mikihiro, et al. *IROS*, 2024.
[5] Chen, Zhiwen, et al. CVPR, 2024.
[6] https://www.eetimes.com/mercedes-applies-neuromorphic-computing-in-ev-concept-car/
[7] Kielty, Paul, et al. *IEEE Access*, 2023.
[8] Chiavazza, S et al. ECCVW, 2025

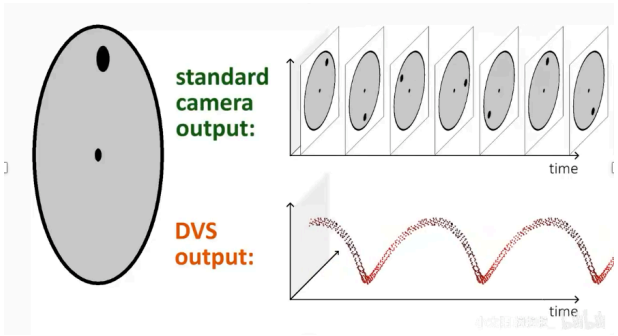
Event Data: Background

What is an Event Camera?

- Novel sensor that measures only **brightness changes (motion)**
- An event is triggered at a **pixel** if

$$\log I(\boldsymbol{x}, t) - \log I(\boldsymbol{x}, t - \Delta t) = \pm C$$





[1] Scaramuzza. Event_Cameras_Tutorial.pdf
[2] Video from https://youtu.be/LauQ6LWTkxM?t=30

Advantages

- Low-latency (~ 1 μs)
- Ultra low power (1 mW vs 1 W)
- Does not suffer from:

Motion blur

Dynamic Range

Disadvantages

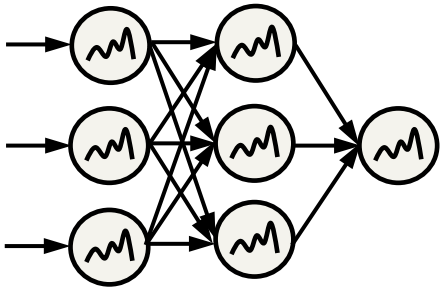
Cannot use traditional vision algorithms (eg. CNN) because:

- Asynchronous data
- No intensity information (only binary intensity changes)
- Difficulty in interpreting event streams into meaningful information → what to do?
  - Naturally aligned with **Spiking neural network (SNNs)**
  - **Requires** novel encoding and new architectures

Yale University

Event Data: Our Research

Event & SNN

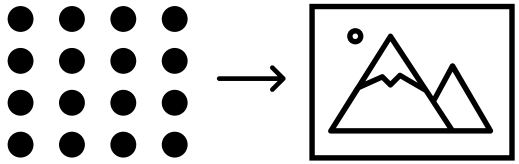


Floating-point based **CNN** with ReLU

VS

Sparse and Spike based **SNN** with LIF

Encoding

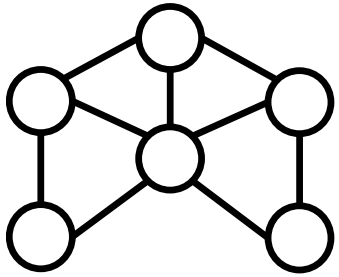


Traditional Event-to-Frame representation

VS

Event-to-Vec

Event-specific Architectures



Transformer

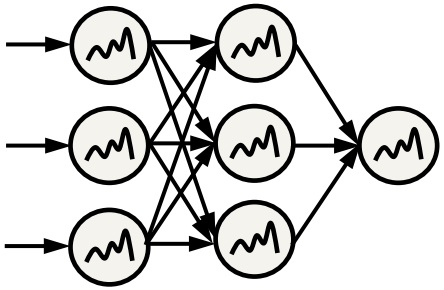
VS

A model which can integrate temporal dynamics

Yale University

Event Data: Our Research

Event & SNN



Floating-point based **CNN** with ReLU

VS

Sparse and Spike based **SNN** with LIF

Encoding

Traditional Event-to-Frame representation

VS

Event-to-Vec

Event-specific Architectures

Transformer

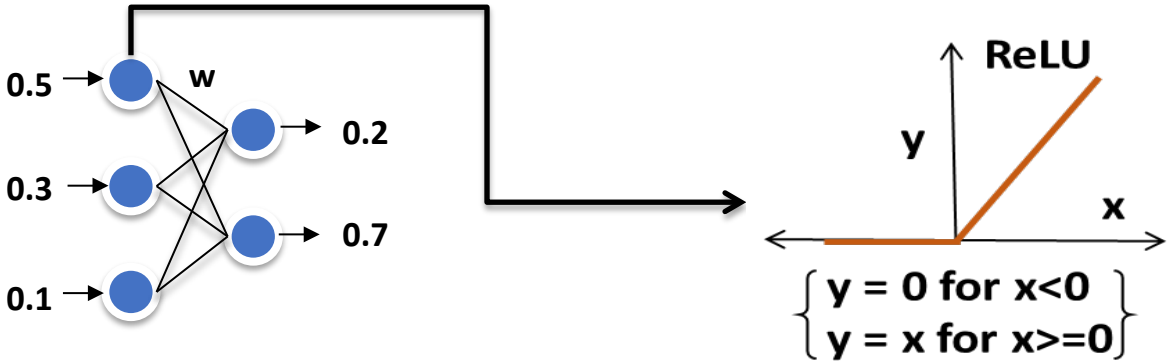
VS

A model which can integrate temporal dynamics

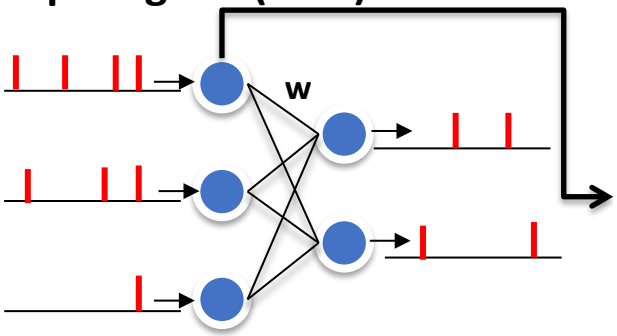
Yale University

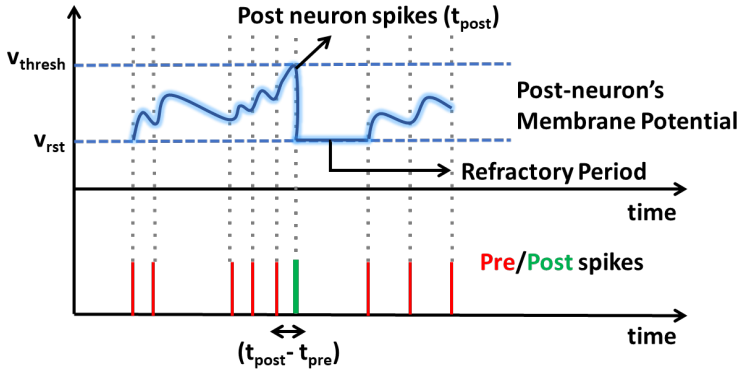
ANN vs. SNN: Fundamental Differences

Artificial NN (ANN)



Spiking NN (SNN)





Leaky-Integrated and Fire (LIF)

Features

(+) High Performance
(+) Various Applications

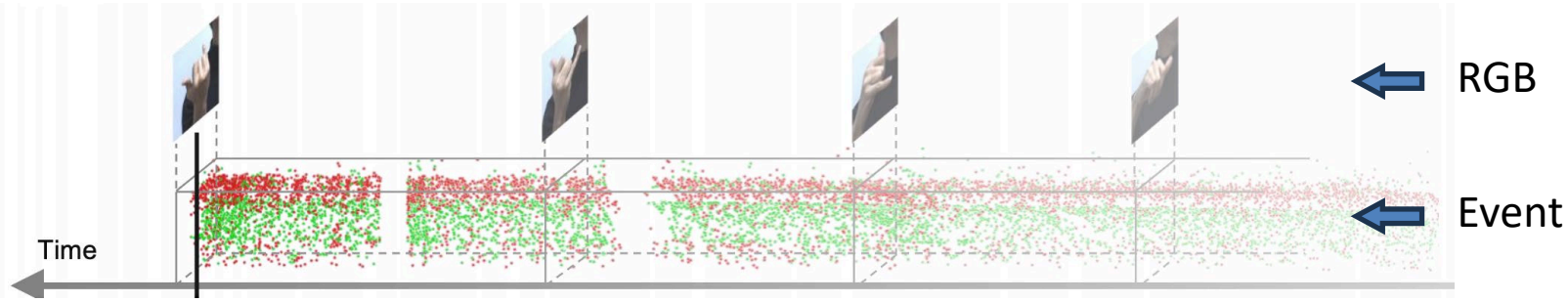
(-) High computational cost with FPs

(+) Low computational cost with spikes (binarized activations)

→ Energy-efficient AI

Yale University

Event Data with Spiking Neural Networks



- **Why SNNs are Good for Event Data:**

Temporal Processing: SNNs process spikes over time

Sparse Activation: Neurons only activate when events occur

Low Latency: Enable low latency processing for real-time robotics applications

Energy Efficiency: Reduce energy consumption through event-driven computation

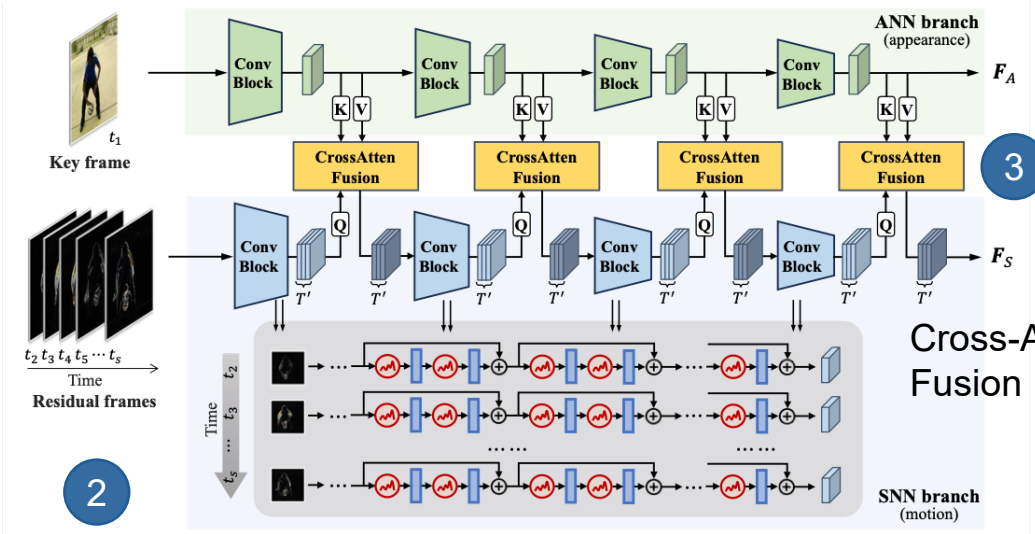
Yale University

Jiang, Jianping, et al. "Complementing event streams and rgb frames for hand mesh reconstruction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

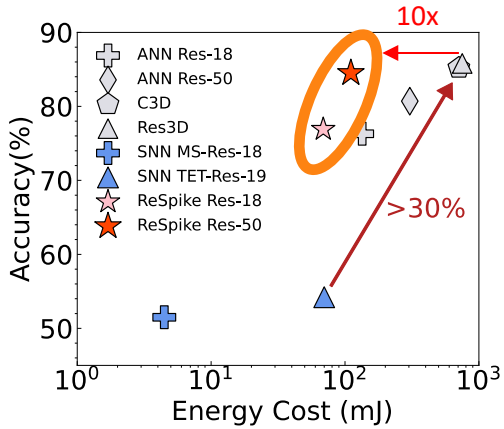
Combine Temporal and Spatial Architectures for RGB+Event Multi-Modal Processing

ReSpike: Residual Frames-based Hybrid Spiking Neural Networks for Efficient Action Recognition

1 ANN branch extracts spatial features from the key frame



Cross-Attention Fusion



UCF-101 / video

SNN branch extracts temporal features from residual frames

Yale University

Xiao, Shiting, et al. "ReSpike: residual frames-based hybrid spiking neural networks for efficient action recognition." Neuromorphic Computing and Engineering (2025).

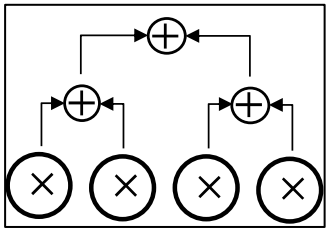
Will Attention Architectures be useful for Events?

- **Bottlenecks of Self-Attention in Standard Transformer**

RGB Transformer

High MAC operations

- QKV Generation
- Projection
- QK^T and SM·V

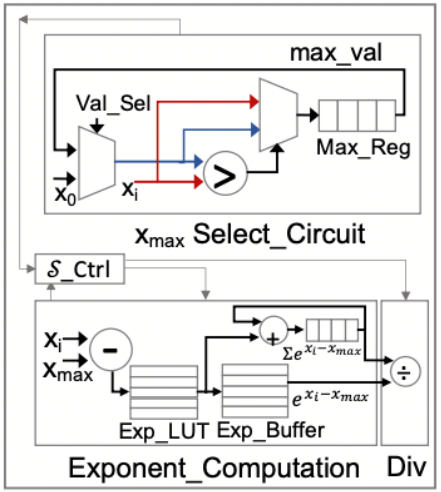


Costly Multipliers

Softmax

$$\frac{e^{x-x_{max}}}{\sum e^{x-x_{max}}}$$

- Sequential Ops
- Complex Ops

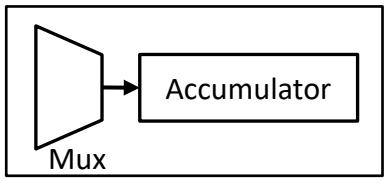


Spiking Event Transformer

$$LIF\left(Query \times Key \times Value \right)$$

Query			Key				Value		
1	0	1	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
1	1	0	1	1	0	1	1	1	0
0	0	1					0	0	1

$$Attn = LIF\{(Q \odot K^T) \odot V\}$$



(+) MAC Operations converted to Mux & Accumulations

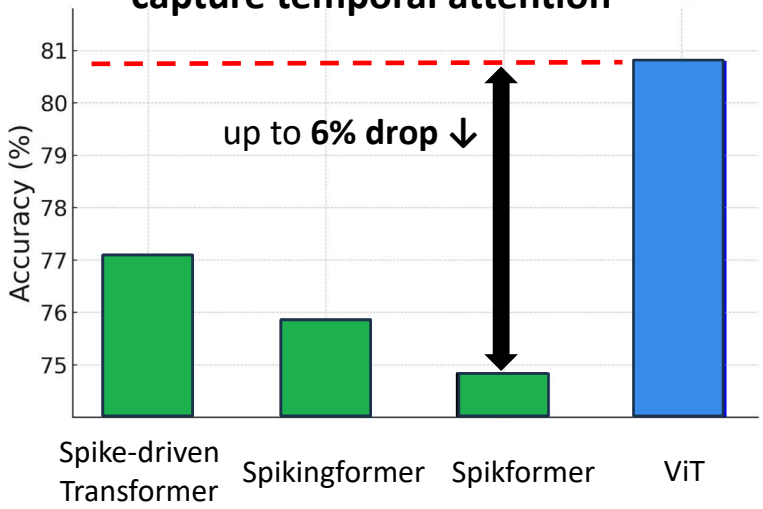
(+) No Softmax Required

Yale University

Moitra, Abhishek, et al. "TReX-Reusing Vision Transformer's Attention for Efficient Xbar-based Computing." *IEEE TETC* 2024.
Lee, Donghyun, et al. "Spiking Transformer with Spatial-Temporal Attention." arXiv:2409.19764 (CVPR 2025).

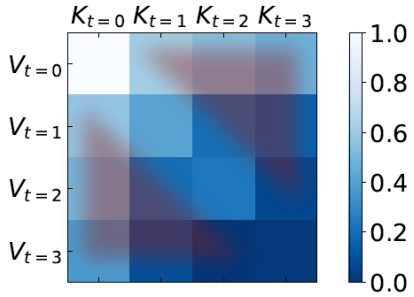
Accuracy Drop with Spiking Transformer

- **Previous architectures did not capture temporal attention**



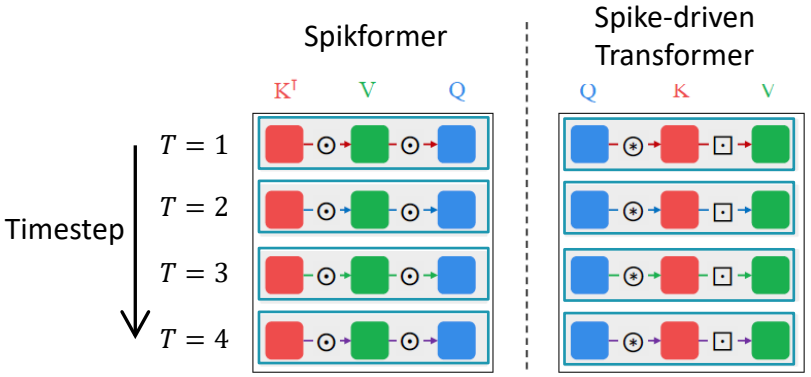
- **Spike Patterns in Self-Attention**

✓ Spike features are various across the timestep



White: high similarity
Blue: low similarity

→ **Q, K, V information are different across the timesteps**



→ **Only spatial correlation**

Zhou, Zhaokun, et al. "Spikformer: When spiking neural network meets transformer." *arXiv preprint arXiv:2209.15425* (2022).

Zhou, Chenlin, et al. "Spikingformer: Spike-driven residual learning for transformer-based spiking neural network." *arXiv preprint arXiv:2304.11954* (2023).

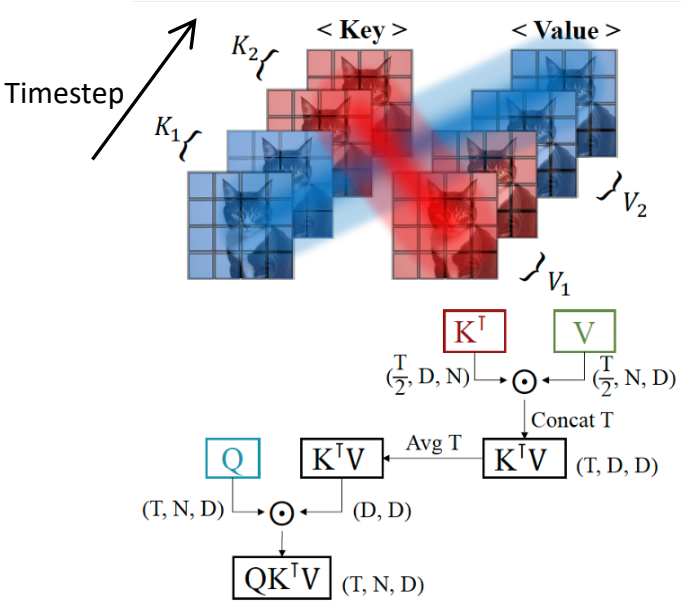
Yao, Man, et al. "Spike-driven transformer." *Advances in neural information processing systems* 36 (2024).

Yale University

Spatio-Temporal Attention (STAtten)

- **Spatio-Temporal Attention (STAtten)**

1) Divide K and V into two groups (K_1, K_2, V_1, V_2)
2) Cross-attention between different timestep



Complexity: $\mathcal{O}(TND^2)$ - Ours (Linear in N)

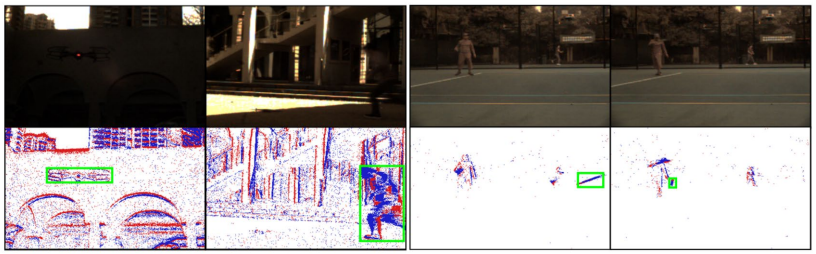
$\mathcal{O}(T^2N^2D)$ - Conventional (Quadratic in N)

Lee, Donghyun, et al. "Spiking Transformer with Spatial-Temporal Attention." arXiv:2409.19764 (CVPR 2025).

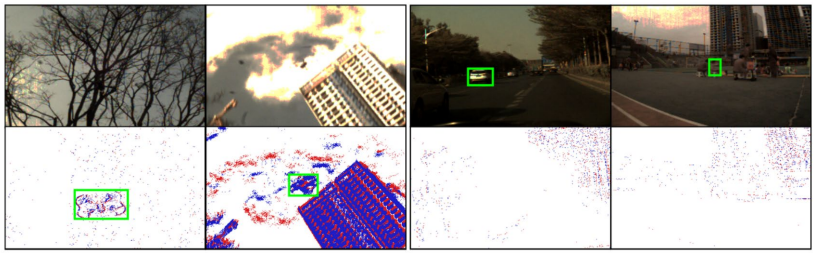
📍 Exhibition Hall D, Poster #315
📅 Sunday, June 14 | 10:30 AM – 12:30 PM

Yale University

Spatio-Temporal Attention (STAtten)



(a) Low illumination (b) High speed



(c) Clutter background (d) Slow motion / static target

DVS dataset

- CIFAR10-DVS, N-Caltech101
- SOTA accuracy with larger accuracy margin

Method	Architecture	Timestep	CIFAR10-DVS (%)	N-Caltech101 (%)
tdBN [8]	ResNet19	10	67.80	-
PLIF [41]	ConvNet	20	74.80	-
Dspike [42]	ResNet18	10	75.40	-
DSR [43]	ResNet18	20	77.27	-
SEW-ResNet [29]	ConvNet	20	74.80	-
TT-SNN [49]	ResNet34	6	-	77.80
NDA [50]	VGG11	10	79.6	78.2
Spikformer [15]	Spikformer-2-256	16	80.9	-
SDT [16]	SDT-2-256	16	80.0	81.80 †
T-STAtten	SDT-2-256	16	**82.2**	**84.36**

Yale University

up to 2.56% improvement

Event Data: Our Research

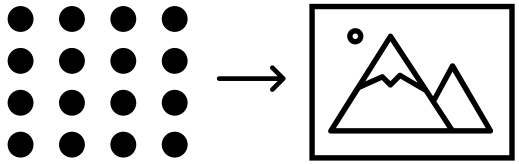
Event & SNN

Floating-point based **CNN** with ReLU

VS

Sparse and Spike based **SNN** with LIF

Encoding



Traditional Event-to-Frame representation

VS

Event-to-Vec

Architectures beyond Transformer

Transformer

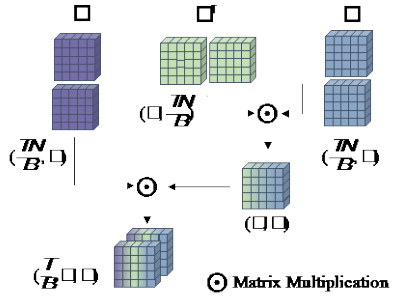
VS

A model which can integrate temporal dynamics

Yale University

Limitations Without Proper Encoding

Spiking Transformer with Spatial-Temporal Attention



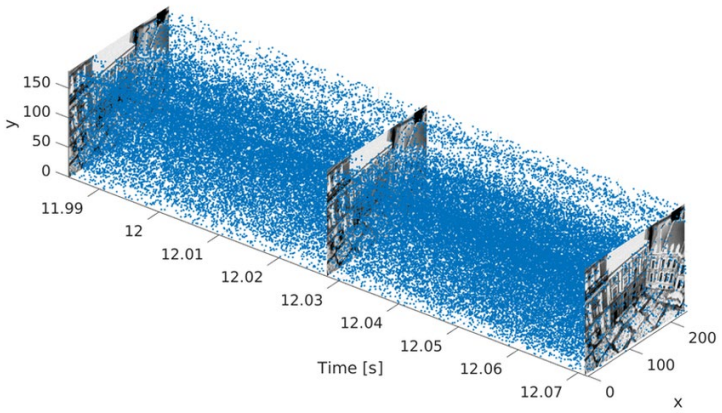
ReSpike: Hybrid Architecture

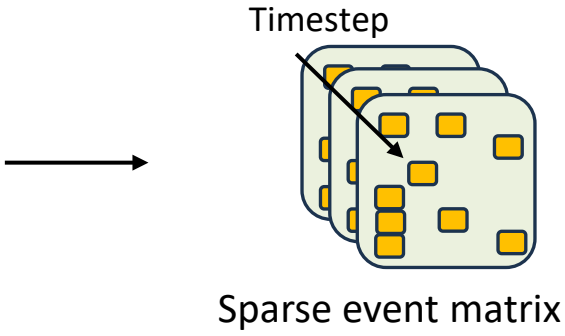
Event frames

Loss of important temporal and spatial event statistics

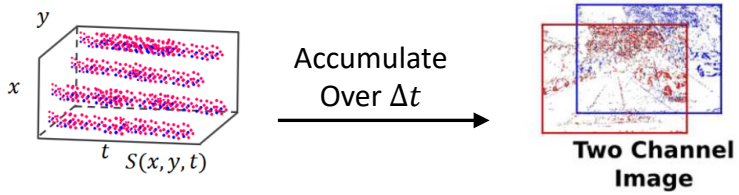
How can we encode the essential structure of event streams?

Fully leveraging Event Sparsity is challenging



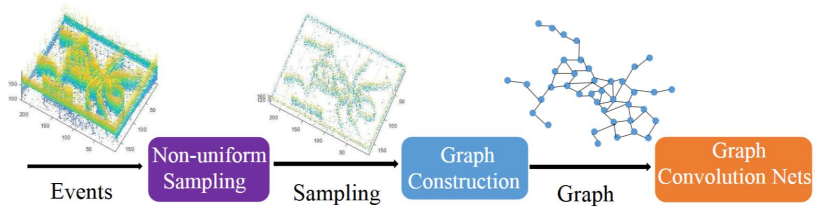


Event-to-frame representation



- Asynchronous-to-synchronous
- Applicable for all neural networks
- Lose the temporal resolution

Event-to-graph representation



- Keep the temporal resolution
- Costly computation to build the graph GNNs
- Can not be deep: "over-smoothing"

Yale University

Inspiration: Word2Vec

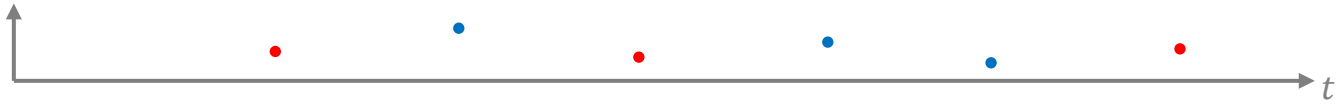
Gehrig et al. End-to-end learning of representations for asynchronous event-based data, ICCV 2019
Schaefer et al. AEGNN: Asynchronous Event-based Graph Neural Networks, CVPR 2022

Event to Vector (Event2Vec)

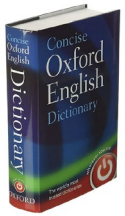
Similarity between events and words in vector spaces

1. Every element is a combination of an index and a position
2. The meaning of an element is determined by its context

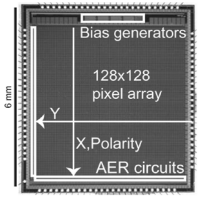
Word	I	am	glad	to	see	you
Index	128000	40	1097	311	1518	499
Position	0	1	2	3	4	5



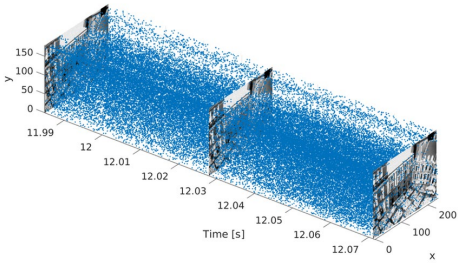
Event	E_0	E_1	E_2	E_3	E_4	E_5
Index	(x_0, y_0, p_o)	(x_1, y_1, p_1)	(x_2, y_2, p_2)	(x_3, y_3, p_3)	(x_4, y_4, p_4)	(x_5, y_5, p_5)
Position	t_0	t_1	t_2	t_3	t_4	t_5



A dictionary contains 600,000 words



A DVS 128 camera contains $2 \times 128 \times 128$ coordinates



Event2vec

$V \in \mathbb{R}^{L \times d}$

$$V = \text{PositionalEncoding}(\text{Embed}(x, y, p), t)$$

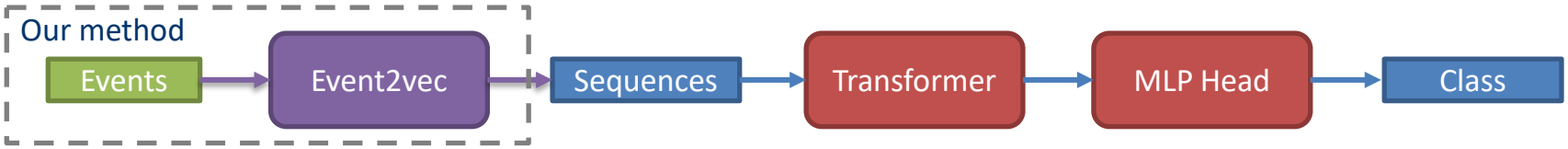
Yale University

Fang, Wei, and Priyadarshini Panda. "Event2Vec: Processing neuromorphic events directly by representations in vector space." *arXiv preprint arXiv:2504.15371* (2025).

Event2vec: Preliminary Results

- Results on ASL-DVS
- We randomly sample 255 events from each event stream

Method	Preprocessing	Accuracy (%)	Model Size (MB)	Epoch
CGG + CNN [1]	To graphs	90.1	19.46	150
GNN + Transformer [2]	To voxel-graphs and images	90.6	220.3	150
Event2vec + Transformer	None	99.68	4.13	64



- ✓ Simplicity (plug-and-play module)
- ✓ Parameter-efficiency
- ✓ Learning-efficiency
- ✓ Speed: 0.23 ms v.s. 16.7ms [2]

Yale University

[1] Graph-based object classification for neuromorphic vision sensing, CVPR 2019
[2] Learning bottleneck transformer for event image-voxel feature fusion based classification, PRCV 2023

Event2Vec: Prospects

Event MLLM

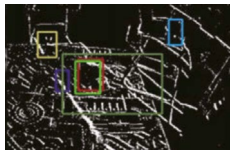
- Event-based multi-modal large language models

"Language prompt"

AR/VR

- AR/VR controller with event camera[1]

- Object tracking[2]



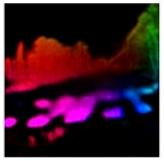
Autonomous Driving

- Keyword spotting [3]

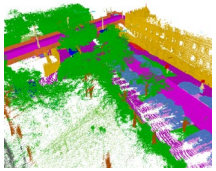
- Driver assistant system [4]

Vision

- Optical flow prediction[5]



- Semantic Segmentation [6]



Future Direction

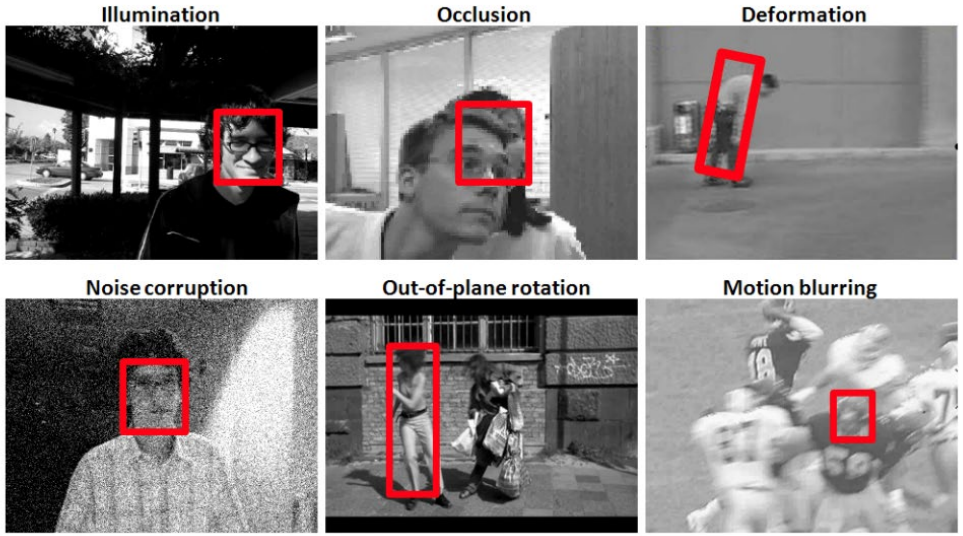
Seamless Integration

[1] Meier, Peter. "AR/VR controller with event camera." U.S. Patent No. 10,845,601. 24 Nov. 2020. (Apple)
[2] Zhang, Jiqing, et al. "Spiking transformers for event-based single object tracking." *CVPR*2022.
[3] https://www.eetimes.com/mercedes-applies-neuromorphic-computing-in-ev-concept-car/
[4] Gehrig, D., Scaramuzza, D. Low-latency automotive vision with event cameras. Nature 629, 1034–1040 (2024).
[5] Chankyu Lee et al., Spike-FlowNet: Event-based Optical Flow Estimation with Energy-Efficient Hybrid Neural Networks (ECCV 2020)
[6] Behley, Jens, et al. "Semantickitti: A dataset for semantic scene understanding of lidar sequences." (CVPR 2019)

Frame/Language Data have Uncertainty

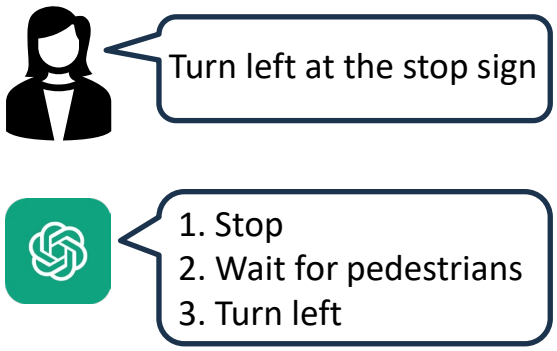
Perception (Visual observation)

- Image artifacts
- Lighting
- Occlusion
- Motion blur
- ……



Decision (Text generation)

- Prompt-image inconsistency
- Failure to capture critical image information
- Failure to incorporate safety rules
- ……



Yale University

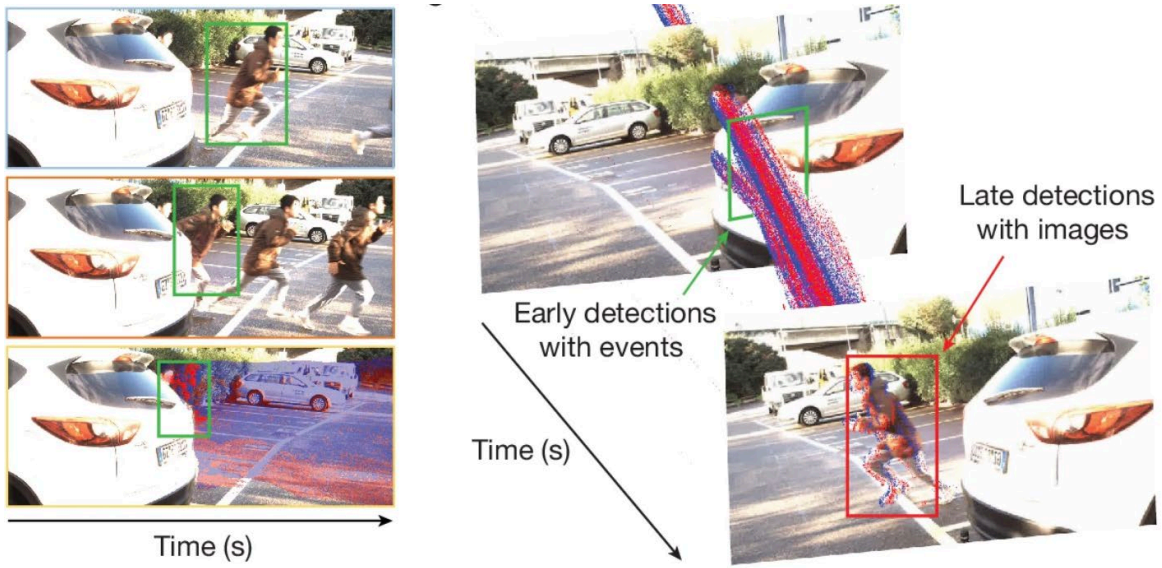
Bhatt, Neel P., et al. “Know Where You‘re Uncertain When Planning with Multimodal Foundation Models: A Formal Framework.” *arXiv preprint (2024)*

Event Data Addresses Uncertainty

Dynamic lighting & high contrast address occlusion

→ Combining a standard 20 FPS camera with an event camera achieved the **equivalent of a 5000 FPS system** in terms of low-latency detection

→ Self-Driving scenario: early detection of motion, enhances safety



Gehrig, D., Scaramuzza, D. Low-latency automotive vision with event cameras. Nature 629, 1034–1040 (2024).

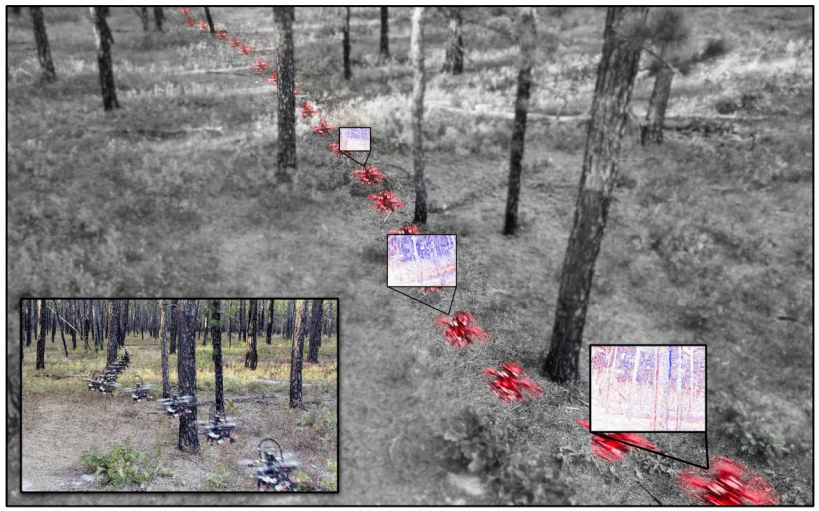
Yale University

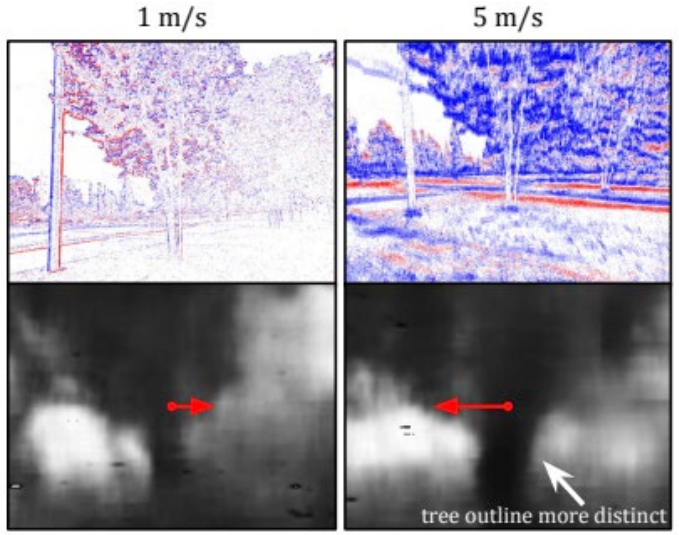
Event Data Addresses Uncertainty

High dynamic range addresses fast action and motion blur

→ **faster speeds make vision tasks easier** for the event camera: abundance of events enable better depth estimation and control

→ Quadrotor navigation scenario: high speed (5m/s) real-world obstacle avoidance





Bhattacharya, Anish, et al. "Monocular event-based vision for obstacle avoidance with a quadrotor." *arXiv preprint* (2024).

Yale University

Event Data: Our Research

Event & SNN

Floating-point based **CNN** with ReLU

VS

Sparse and Spike based **SNN** with LIF

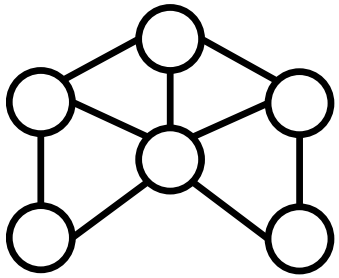
Encoding

Traditional Event-to-Frame representation

VS

Event-to-Vec

Event-specific architecture



Transformer

VS

A model which can integrate temporal dynamics

Yale University

Event-based Architecture: What's needed?

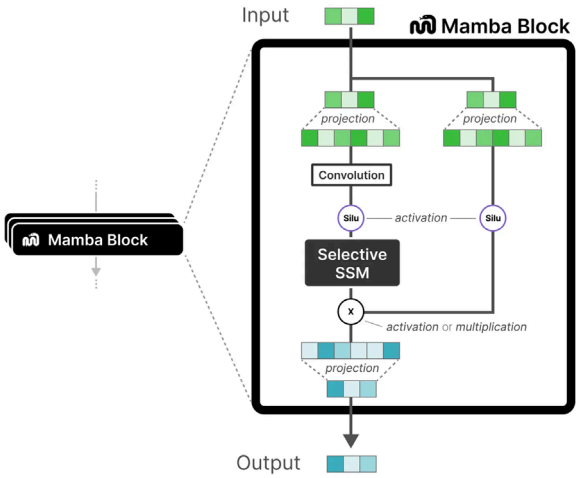
- **Escaping overhead from self-attention and KV cache**

  → **Various linear attention**: Performer[1], Linear Transformer[2]

  ✓ Reducing complexity and handling long sequences

  → **Event dataset:** Temporal processing, Sparse dynamics

  → Beyond self-attention: **State-Space Model (SSM) or Mamba**



- ✓ **Modeling of memory, retention, selective forgetting**
- ✓ **Linear time complexity**
- ✓ **Long sequence handling**

300M Parameter	Commonsense Reasoning (%)	Throughput (token/sec)	Cache size (MB)
Transformer (Llama)	**44.08**	721.1	414.7
Mamba	42.98	**4720.8**	**1.9**

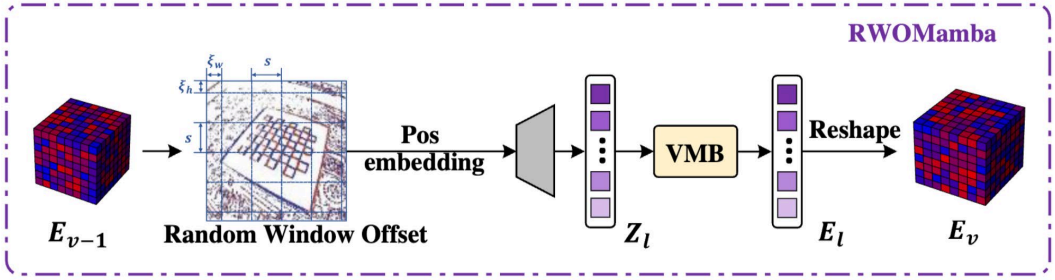
✓ Becoming a foundation model for several downstream tasks

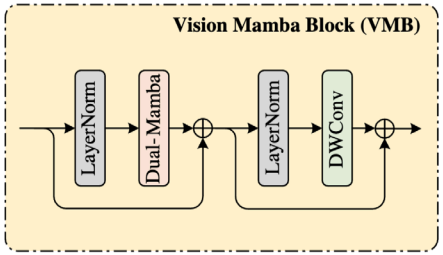
Yale University

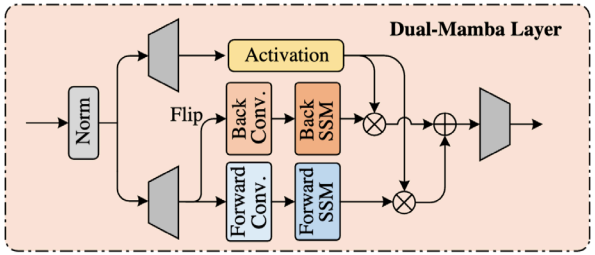
[1] Choromanski, Krzysztof, et al. "Rethinking attention with performers." *arXiv preprint arXiv:2009.14794* (2020).
[2] Katharopoulos, Angelos, et al. "Transformers are rnns: Fast autoregressive transformers with linear attention." *International conference on machine learning*. PMLR, 2020.
Dong, Xin, et al. "Hymba: A hybrid-head architecture for small language models." *arXiv preprint arXiv:2411.13676* (2024).

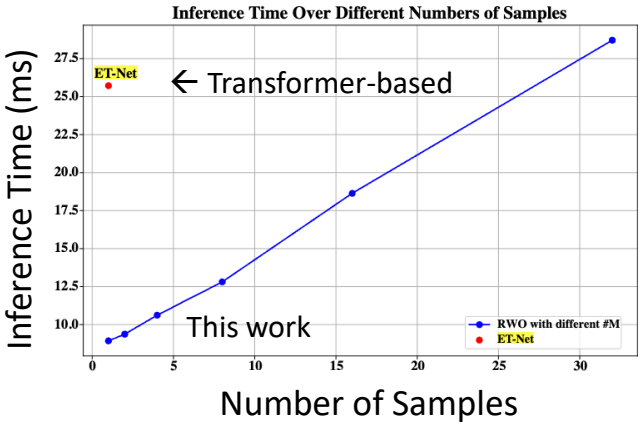
Event-based Architecture: What's needed?

- Existing works
  → EventMamba: Integration of recurrent dynamics with Mamba for event dataset









- Pros: based on given dataset, Mamba architecture is very efficient and effective this sparse and long sequence dataset
- Cons: Adapting rapidly to new information is still limited

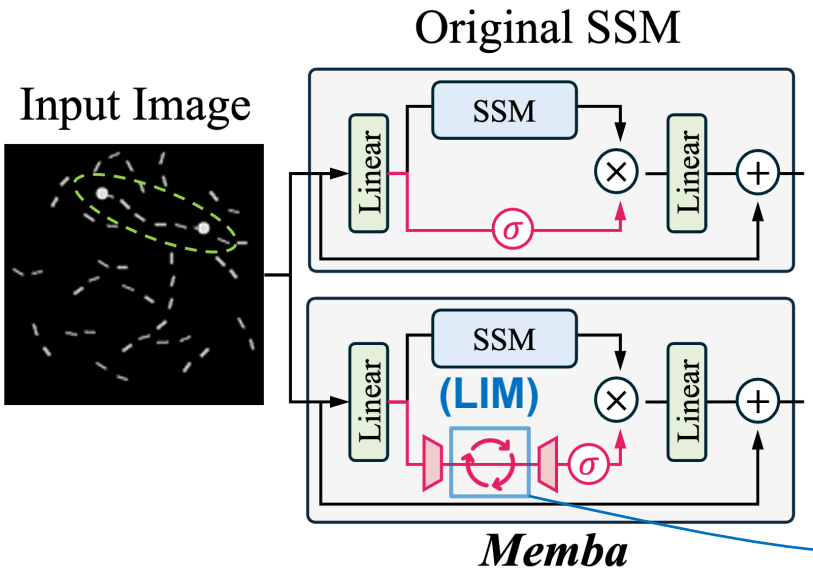
✓ *How to effectively and efficiently adapt the Mamba to new environment?*

Yale University

Ge, Chengjie, et al. "EventMamba: Enhancing Spatio-Temporal Locality with State Space Models for Event-Based Video Reconstruction." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. No. 3. 2025.

Memba: Membrane-driven Parameter-Efficient Fine-Tuning for Mamba (under progress)

- **Enhancing SSM gate with recurrent hidden states**
- **Leaky Integate Membrane (LIM):** Inspired by Spiking Neural Networks



Leaky Integrate Membrane (LIM)

$$\mathbf{u}[i+1]^l = r(\tau \mathbf{u}[i]^l + \mathbf{W}^l X[i]),$$

$$r(x) = \begin{cases} 0 & \text{if } x > V_{th}, \\ x & \text{otherwise} \end{cases},$$

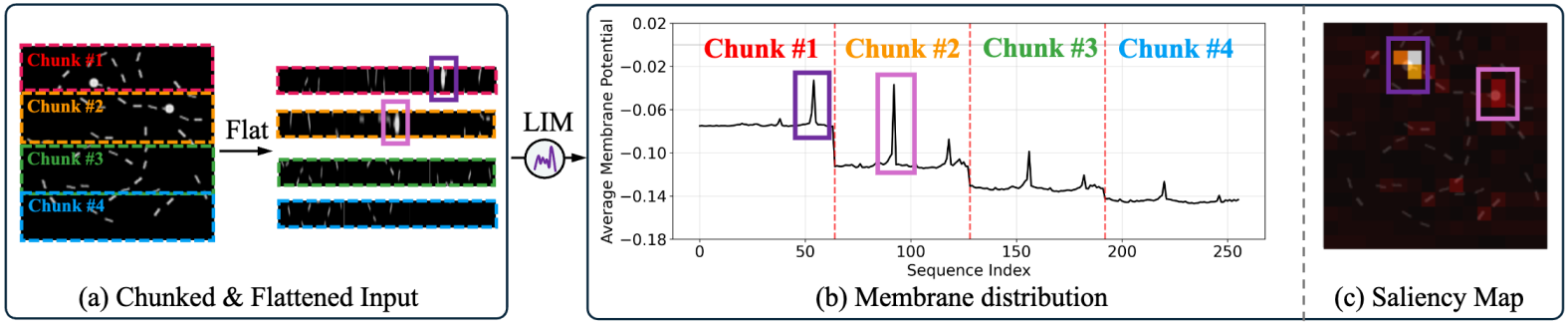
u: membrane
τ: leaky factor
V_{th}: threshold

✓ Inspired by LIF neuron in SNN, we use **membrane potential as an output**, not spike outputs
✓ Through LIM, we bring temporal adaptation during fine-tuning

Yale University

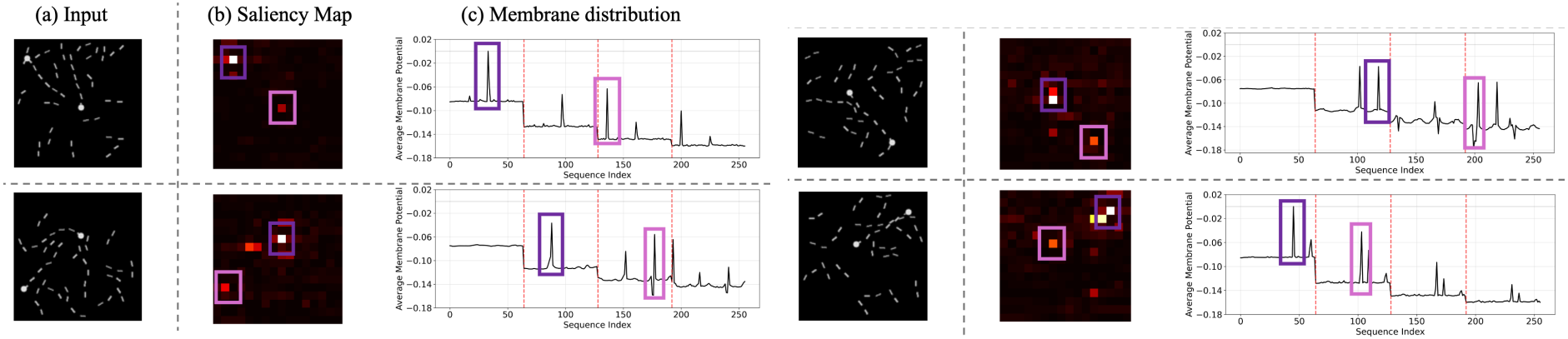
Memba: Membrane-driven Parameter-Efficient Fine-Tuning for Mamba (under progress)

- **Membrane driven Temporal Processing**



✓ Capturing critical features, shows **high peak** in membrane potential (purple & pink boxes)

✓ Graudal decrease in membrane potential, indicating progressive **forgetting memory**



- Accuracy on PathFinder dataset:

Original SSM (S4)	94.20 %
Memba	**97.21 %**

Yale University

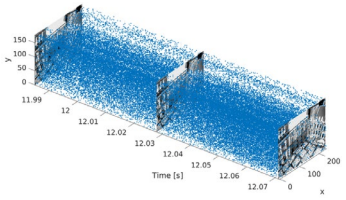
Event2Vec + Ma(e)mba

- **Most of the Event data for Mamba is generally encoded to frames**

➢ **Event2Vec**

✓ Transform asynchronous spike events → vector token
✓ Learn a spatio-temporal embedding simultaneously

+ Ma(e)mba

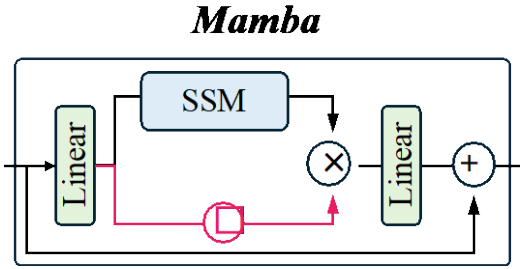


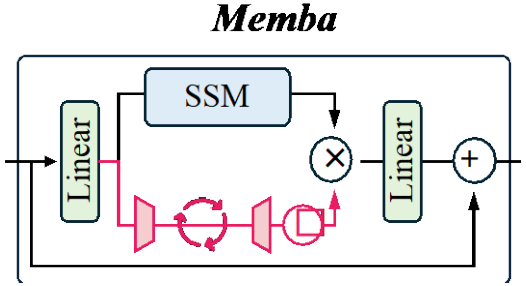
Event2vec

$$V = \text{PositionalEncoding}(\text{Embed}(x, y, p), t)$$

Training

Fine-tuning



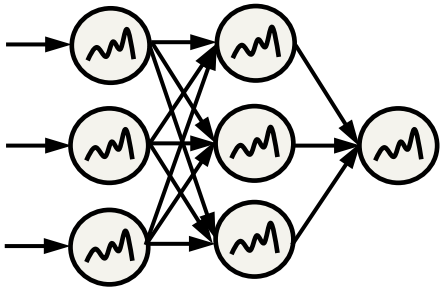


?

Yale University

Summary

Event & SNN

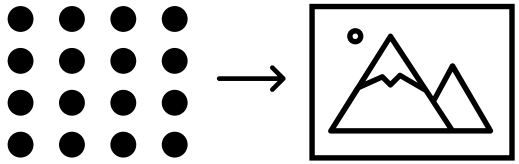


Floating-point based **CNN** with ReLU

VS

Sparse and Spike based **SNN** with LIF

Encoding

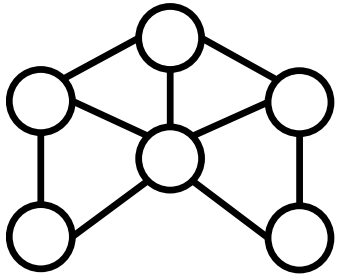


Traditional Event-to-Frame representation

VS

Event-to-Vec

Event-specific Architectures



Transformer

VS

A model which can integrate temporal dynamics