Falcon: An Event-Driven Object Tracking SoC with Activation-Skip and Weight-Compression MRAM PIM and Heterogenous Q/K/V Read-Only-Once Attention Flow

Wenao Xie*, Haoyang Sang*, Zhamaliddin Kalzhan*, Jingu Lee*, Jerald Yoo†, Kangho Lee‡, Hoi-Jun Yoo*
**KAIST, Daejeon, Republic of Korea*
†*Seoul National University, Seoul, Republic of Korea*
‡*Samsung Electronics, Yongin, Republic of Korea*
Email: xwa15@kaist.ac.kr

***Abstract*—This paper presents Falcon, an event-driven object tracking SoC integrating heterogeneous MRAM computing blocks with three features: 1) An activation-skip and weight-compression MRAM processing-in-memory (PIM) for CNN encoders with 2.44× efficiency and 61.9% latency reduction. 2) A Q/K/V read-only-once attention flow featuring the column-priority generation on near memory computing (NMC), the mixed-product matrix multiplication on SIMD, and the precision-enhanced SoftMax on LUT, achieving buffer read and latency reductions of 46% and 30%, respectively; 3) A two-stage event-detection (ED) system with system power and latency reductions of 63.5% and 68.8%, respectively. Fabricated in 28nm FDSOI technology, the 12.96 mm^2 Falcon demonstrates state-of-the-art macro and system energy efficiencies of 750.18 TOPS/W (1b-1b-3b) and 16.34 TOPS/W (INT8 CNN + FP8 attention), respectively, and achieves the highest hybrid object tracking speed of 50 fps at 1.15V/200MHz.**

***Index Terms*—MRAM, processing-in-memory, object tracking, edge SoC, heterogeneous, attention, event-driven**

I. Introduction

Hybrid object tracking combines RGB and event cameras to enhance tracking accuracy and speed on extreme scenarios, such as motion blur, high dynamic range and fast moving [1], [2]. However, its heavy workload limits its employment on edge SoCs. To tackle this, an event-driven edge SoC with MRAM [3] was introduced to reduce the system's stand-by power. Also, sparsity-aware PIM designs [4], [5] were employed to improve PIM's energy efficiency. However, there remain three challenges: 1) Similar weight and activation bit columns cause duplicate PIM operations; 2) Attention layer is under-utilized on PIM, resulting in shortage of the bandwidth and capacity of the on-chip MRAM; 3) Even non-event frames and patches incur redundant tracking, as shown in Fig. 1.

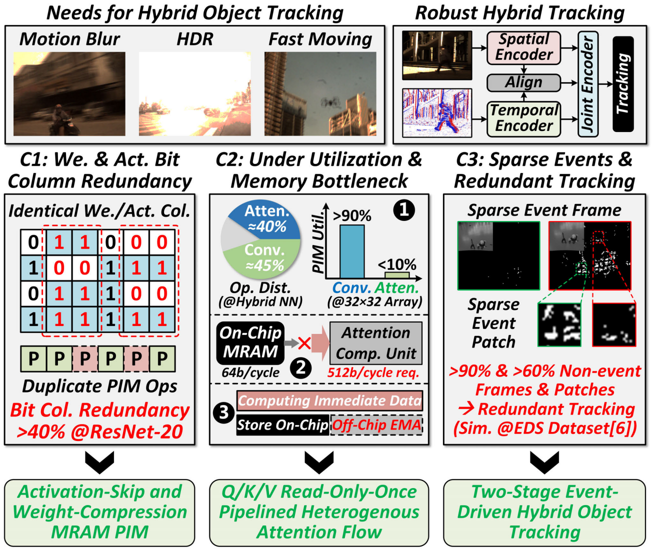


Fig. 1. Hybrid object tracking in edge devices: introduction and challenges.

II. Proposed System

A. Overall Architecture

The proposed Falcon SoC is composed of a RISC-V core, a 256KB global SRAM, a SIMD block, and an event block, as shown in Fig. 2. Also, Falcon contains a MRAM PIM block with 32 MRAM PIM cores, each containing 9 PIM banks and 9 activation drivers. Each PIM bank includes 16 PIM macros. Besides, Falcon contains an eMRAM NMC block with an 8Mb eMRAM and 8 NMC cores, and a MRAM LUT block with 10 MRAM LUT cores. The data router, equipped with 4 different block buffers, manages data flow among these computing blocks. In addition, Falcon integrates an image pre-processing block to align inputs from both event and RGB modalities. This includes temporal synchronization between RGB frames and event streams, spatial alignment using camera calibration, and resizing both modalities to a unified resolution, facilitating effective event-driven object tracking.

B. Activation-Skip and Weight-Compression MRAM PIM

Falcon optimizes multi-channel activation and weight MAC operations on the MRAM PIM macro and activation driver (Fig. 3). Identical adjacent weight bit columns (BCs) optimized by off-line BC similarity tuning, are compressed,

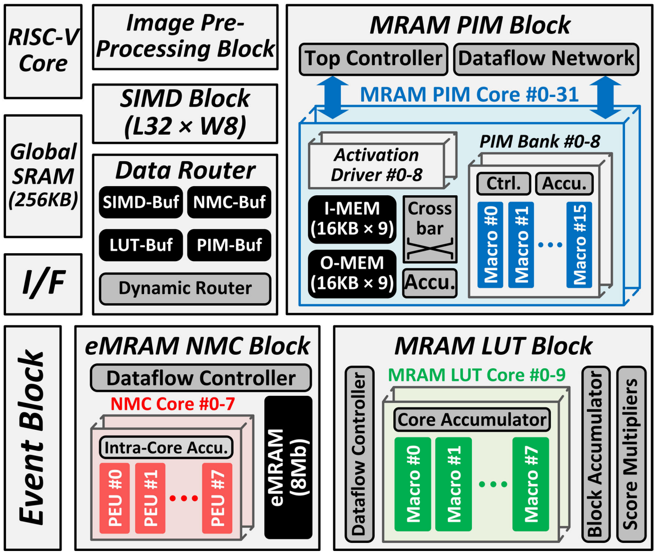
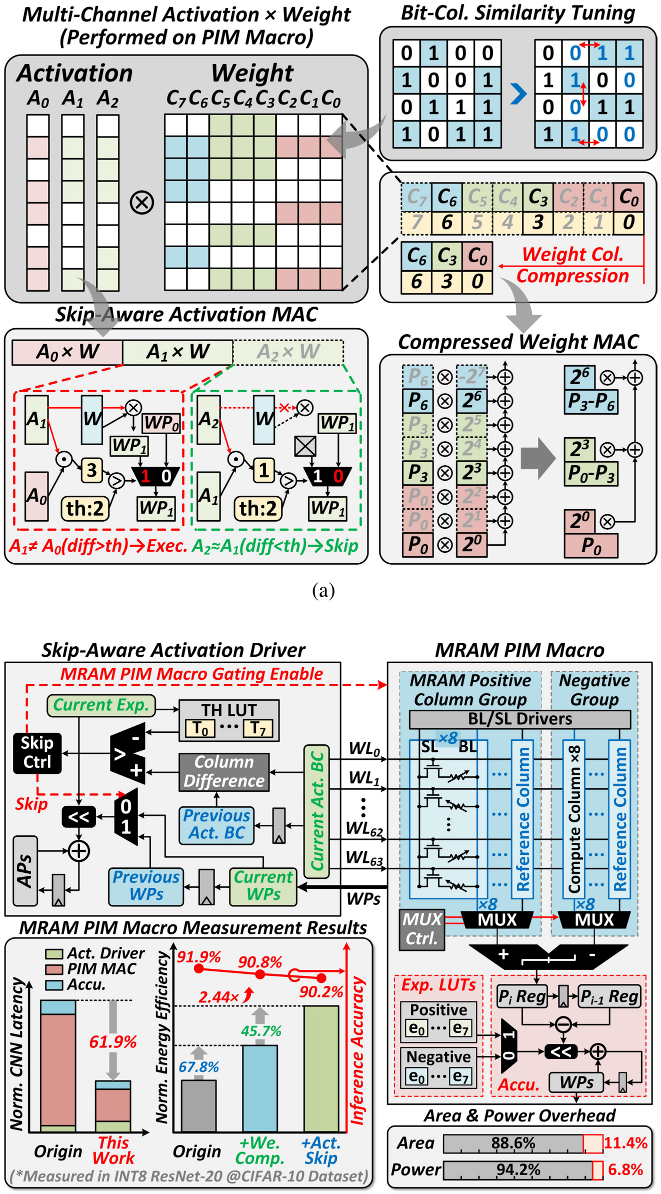


Fig. 2. The overall architecture of the proposed system.



(a)

(b)

Fig. 3. Weight compression and activation-skip MRAM PIM MAC: (a) motivation and methods. (b) architectures and measurement results.

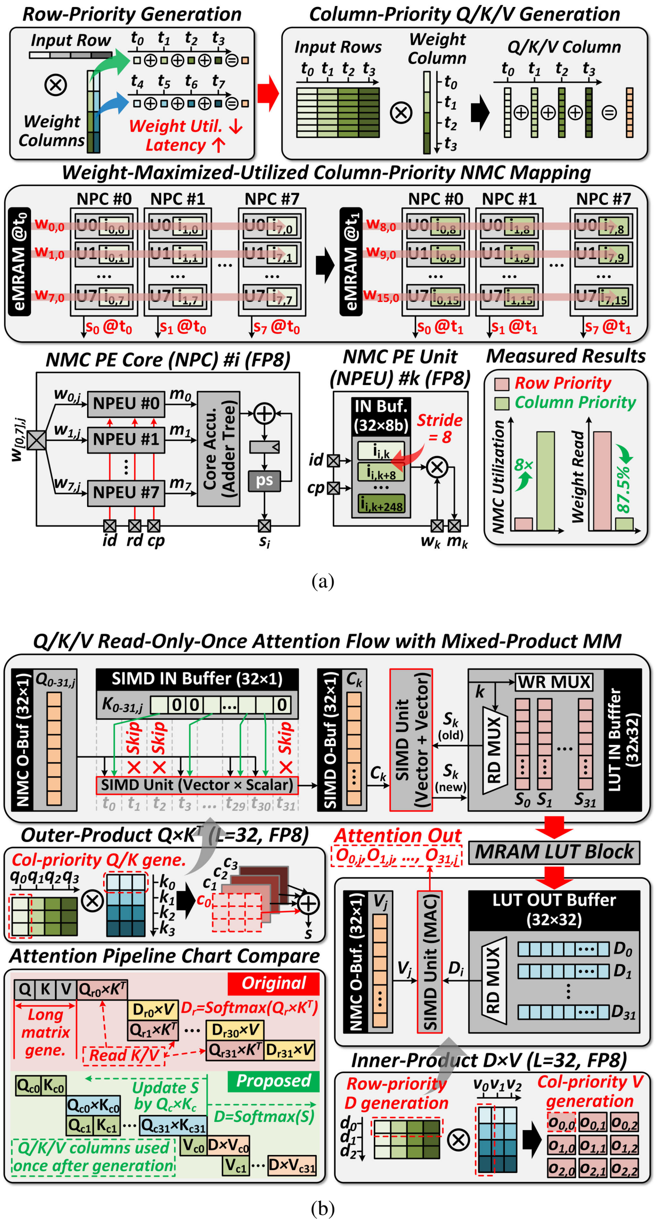
allowing the merging of consecutive shifts of the same partial sum (P_i) to reduce the number of additions and shifts at the weight-level MAC. Besides, if the dynamic current activation BC (A_i) differs minimally ($< threshold$) from its previous one (A_{i-1}), multiplication with W could be skipped by reusing the previous weight-level partial sums (WPs). The MRAM PIM macro performs time-multiplexed MAC with compressed weights utilizing a shared accumulation circuit and exponent LUTs. Also, the skip-aware activation driver buffers the activation BCs and WPs, where the activation-level partial sums (APs) sum up either the previous or current WPs based on whether the current activation BC is skipped or not. Despite an area increase of 11.4% and a power overhead of 6.8%, the proposed design significantly reduces latency by 61.9% and enhances energy efficiency by 2.44× for CNN workloads compared to the original non-optimized PIM macro, with only a 1.7% accuracy loss in ResNet-20 on CIFAR-10 dataset.

C. Q/K/V read-only-once attention flow

During the Q/K/V generation process (Fig. 4a), the NMC PE unit stores inputs at a stride of 8, and the NMC PE core aggregates the multiplication results from 8 units. To maximize the limited eMRAM weight read bandwidth, the proposed column-priority (COP) matrix multiplication (MM) strategy multiplies 8 input rows with a shared weight column, achieving 8× utilization and 87.5% less weight read. During the Q/K/V read-only-once attention flow (Fig. 4b), Given that Q and K are both COP generated, Q×K^T adopts the outer-product MM and updates the attention score matrix (S) with a multi-mode SIMD unit. Once S is finalized, the MRAM LUT Block performs the SoftMax operation and generates the attention distribution matrix (D) in a row-priority (ROP) style. Thus, the inner-product MM is employed to D×V (COP generated). Compared to the original attention flow with row-wise D_r generation and repeated K/V reading, the proposed matrix-wise D generation-based pipeline utilizes the Q/K/V columns only once after their generation. The MRAM LUT block (Fig. 5) performs SoftMax with enhanced precision from FP8 to FP16. It obtains the natural exponential values (ex, FP8) of the D rows from the 2T2R-cell-based MRAM LUT macros [3], as well as their sum's reciprocal (FP16). Measurements are conducted on one attention layer head with a length (L) of 32 and head (H) of 8. Even as the feature dimensions increase, the head's required buffer in the proposed design remains at 2KB, achieving about 46% and 30% reductions in head buffer read and head latency across all feature dimensions, respectively.

D. Two-Stage Event-Detection (ED) System

Within the proposed two-stage ED system (Fig. 6), the event frame processing block (EFPB) receives a gray image



(a)

(b)

Fig. 4. (a) Q/K/V generation workflow. (b) Full attention layer workflow.

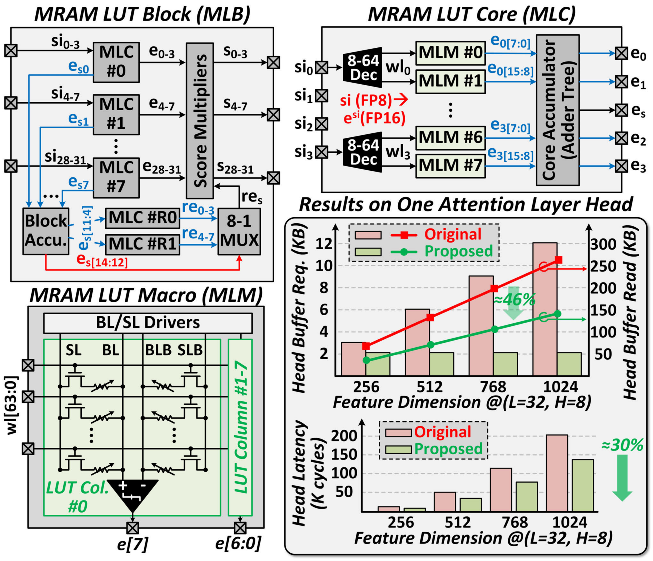


Fig. 5. MRAM look-up-table (LUT) block, core, and macro: architectures and measurement results.

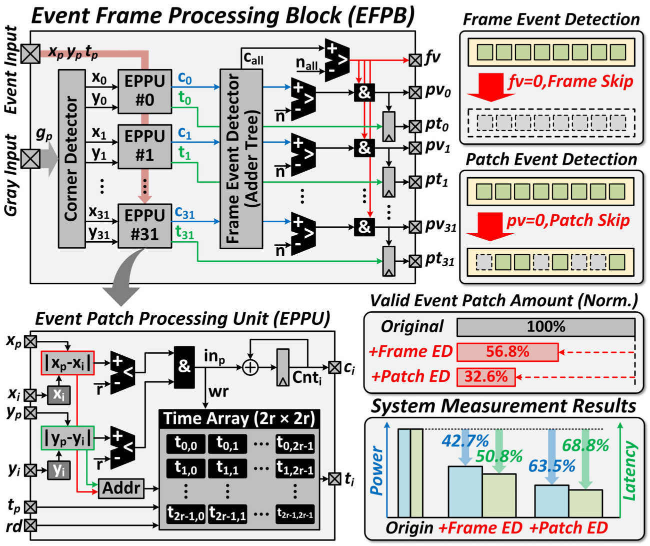
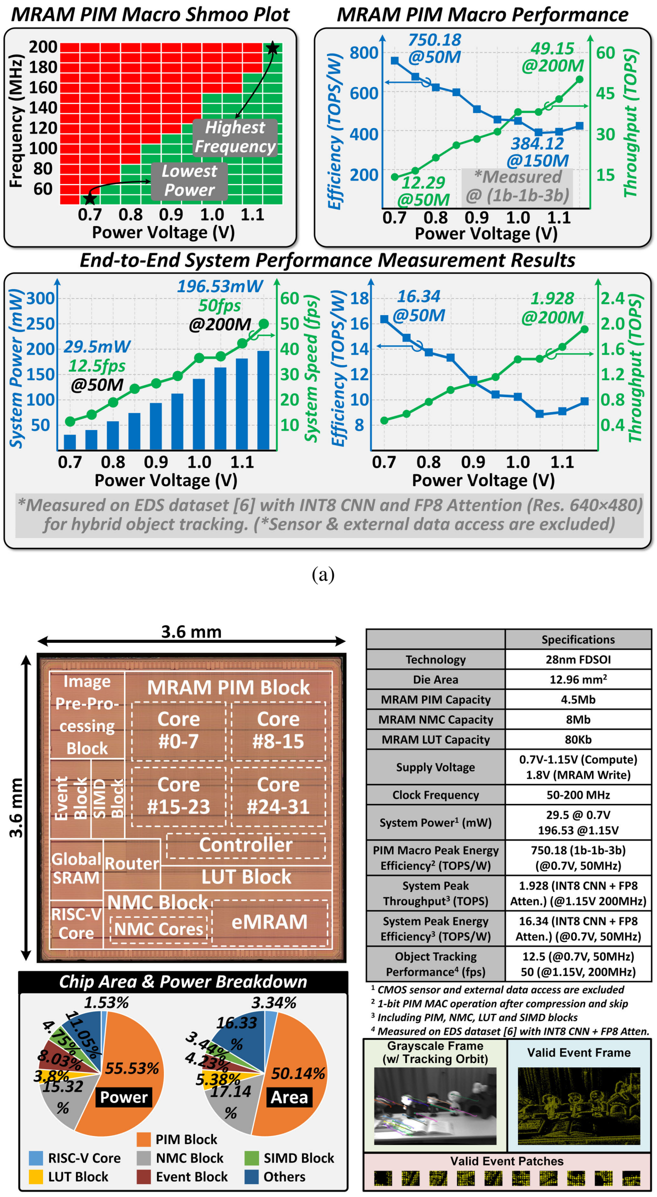


Fig. 6. Event frame processing block and event patch processing unit: architectures and measurement results.

input (g_p) and performs corner detection to identify 32 corner coordinates (x_i, y_i), which are designated as patch centers in corresponding event patch processing units (EPPUs). The EPPU processes event image input (x_p, y_p, t_p), storing tp into the time array if both the x- and y-offsets from the center (x_i, y_i) fall within a defined patch radius (r). Besides, the count of valid event pixels (c_i) recorded in each EPPU is aggregated in the EFPB, where it is compared with threshold (n_{all}) to determine if the event frame is valid or not (fv). At the same time, each c_i is compared with another threshold (n) to determine the validity of event patches (pv). The outputs fv and pvs are used to manage the frame and patch ED, avoiding invalid event frames and patches. The measurement results show a reduction of 67.4% in the number of valid event patches, with overall system power and latency reductions of 63.5% and 68.8%, respectively.

III. Conclusion

Fig. 7a shows the SoC's measurement results. The MRAM PIM macro's shmoo plot shows that it achieves the lowest power and highest energy efficiency of 750.18TOPS/W (1b-1b-3b) at 0.7V/50MHz. It also reaches the highest frequency and throughput of 49.15TOPS (1b-1b-3b) at 1.15V/200MHz. Also, in end-to-end hybrid object tracking using 8-bit activation and weight on the EDS dataset [6] at 640×480 resolution, Falcon achieves the highest energy efficiency of 16.34 TOPS/W and the highest throughput of 1.928 TOPS. Falcon



	Specifications
Technology	28nm FDSOI
Die Area	12.96 mm^2
MRAM PIM Capacity	4.5Mb
MRAM NMC Capacity	8Mb
MRAM LUT Capacity	80Kb
Supply Voltage	0.7V-1.15V (Compute) 1.8V (MRAM Write)
Clock Frequency	50-200 MHz
System Power[1] (mW)	29.5 @ 0.7V 196.53 @1.15V
PIM Macro Peak Energy Efficiency[2] (TOPS/W)	750.18 (1b-1b-3b) (@0.7V, 50MHz)
System Peak Throughput[3] (TOPS)	1.928 (INT8 CNN + FP8 Atten.) (@1.15V 200MHz)
System Peak Energy Efficiency[3] (TOPS/W)	16.34 (INT8 CNN + FP8 Atten.) (@0.7V, 50MHz)
Object Tracking Performance[4] (fps)	12.5 (@0.7V, 50MHz) 50 (@1.15V, 200MHz)

[1] *CMOS sensor and external data access are excluded*
[2] *1-bit PIM MAC operation after compression and skip*
[3] *Including PIM, NMC, LUT and SIMD blocks*
[4] *Measured on EDS dataset [6] with INT8 CNN + FP8 Atten.*

Fig. 7. (a) Measurement results. (b) Chip die photo, specifications, area and power breakdown, and event detection results.

also supports object tracking speeds up to 50fps at 200MHz and maintains 12.5fps at its lowest power consumption of 29.5mW. Falcon is fabricated in 28nm FDSOI technology and occupies a die area of 12.96mm^2. The chip die photo, specifications, area and power breakdown, and event detection results are provided in Fig. 7b. Compared to other edge SoCs [7]–[11] in Table. I, Falcon achieves state-of-the-art macro and system energy efficiencies, as well as hybrid object tracking speed.

Acknowledgments

This work was supported by IITP grant funded by Korea government (MSIT) (IITP-2024-RS-2023-00256472, Graduate School of AI Semiconductor; 2022-0-01170, PIM Semiconductor Design Research Center).

TABLE I
Comparison with state-of-the-art edge SoCs

	VLSI'23 [3]	VLSI'22 [7]	ISSCC'23 [8]	ISSCC'24 [9]	VLSI'24 [10]	VLSI'24 [11]	This Work
Technology (nm)	28	22	40	40	40	12	28
Area (mm^2)	9.72	8.76	20.25	20.25	30	0.5	12.96
Processing Type	PIM	NMC	CIM+NMC	NMC	NMC	CIM+NMC	PIM/NMC/LUT
Computing Domain	Analog	Digital	Hybrid	Digital	Digital	Analog	Hybrid
PIM/CIM Capacity	128Kb MRAM	N/A	1.25MB RRAM	N/A	N/A	186KB SRAM	4.5Mb MRAM
NMC Capacity	N/A	2MB MRAM	1.25MB SRAM	5MB RRAM	1.8MB SRAM	64KB SRAM	8Mb eMRAM
LUT Memory	N/A	N/A	N/A	N/A	N/A	N/A	80Kb MRAM
Voltage (V)	1	0.5-1.0	0.9	0.8-1.1	0.6-1.1	0.64	0.7-1.15
Frequency (MHz)	200	0.056-190	100	80-210	350	600	50-200
Peak Macro Throughput (TOPS)	-	-	14.74 (1b)	-	2.25 (8b)	-	49.15 (1b)
Peak Macro Energy Efficiency (TOPS/W)	709.3 (1b-1b)	-	-	-	29.45 (8b)	86 (6b)	750.18 (1b)
System Power (mW)	7.3mW-12.16mW	468uW-158mW	4.6mW-21.3mW	110uW-603.9mW	2.1mW	No	29.5mW-196.53mW
System Throughput (TOPS)	-	0.511 (8b)	-	0.1-0.269 (8b)	-	3.4 (6b)	1.928 (8b)
System Energy Efficiency (TOPS/W)	-	12.1 (8b)	73.53 (1b)	0.84-1.14 (8b)	-	46 (6b)	16.34 (8b)
Application	CNN Inference	CNN, Robot Navigation	CNN+SNN Targe Tracking	CNN, Micro Surveillance	CNN	CNN	CNN+Atten. Object Tracking

References

[1] N. Messikommer et al., "Data-driven feature tracking for event cameras," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5642–5651.

[2] X. Wang et al., "Visevent: Reliable object tracking via collaboration of frame and event flows," IEEE Transactions on Cybernetics, vol. 54, no. 3, pp. 1997–2010, 2023.

[3] W. Xie et al., "A 709.3 TOPS/W event-driven smart vision SoC with high-linearity and reconfigurable MRAM PIM," in 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2023, pp. 1–2.

[4] W.-H. Huang et al., "A nonvolatile Al-edge processor with 4MB SLC-MLC hybrid-mode ReRAM compute-in-memory macro and 51.4-251TOPS/W," in 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, pp. 15–17.

[5] T.-H. Wen et al., "34.8 a 22nm 16mb floating-point reram compute-in-memory macro with 31.2 tflops/w for ai edge devices," in 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, vol. 67, pp. 580–582.

[6] J. Hidalgo-Carrió et al., "Event-aided direct sparse odometry," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5781–5790.

[7] Q. Zhang et al., "A 22nm 3.5 TOPS/W flexible micro-robotic vision SoC with 2MB eMRAM for fully-on-chip intelligence," in 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2022, pp. 72–73.

[8] M. Chang et al., "A 73.53 TOPS/W 14.74 TOPS heterogeneous RRAM in-memory and SRAM near-memory SoC for hybrid frame and event-based target tracking," in 2023 IEEE International Solid-State Circuits Conference (ISSCC), 2023, pp. 426–428.

[9] S. D. Spetalnick et al., "30.1 A 40nm VLIW edge accelerator with 5MB of 0.256 pJ/b RRAM and a localization solver for bristle robot surveillance," in 2024 IEEE International Solid-State Circuits Conference (ISSCC), 2024, vol. 67, pp. 482–484.

[10] A. Gupta et al., "CogniVision: End-to-End SoC for Always-on Smart Vision with mW Power in 40nm," in 2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2024, pp. 1–2.

[11] J. Yang et al., "A 278-514M event/s ADC-less stochastic compute-in-memory convolution accelerator for event camera," in 2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2024, pp. 1–2.