A 22.31 TFLOPS/W 73 Mb ReRAM-based ROI-Aware Accelerator for Next-generation AR/VR

Pufan Xu[1], Peng Yao[1]*, Jianing Li[1], Tianhao Yang[1], Jiaming Li[1], Yuepeng Jiang[1], Rui Li[1], Lifan Xie[1], Songrui Zhang[1], Zhenqi Hao[1], Jianxing Liao[2], Wei Yang[2], Leibin Ni[2], Taiwei Chiu[3], Qingtian Zhang[1], Jianshi Tang[1], He Qian[1], Bin Gao[1], Huaqiang Wu[1]

[1]School of Integrated Circuits, BNRist, Tsinghua University, Beijing, China [2]ACS Lab, Huawei Technologies Co., Ltd., Shenzhen, China
[3]Xiamen Industrial Technology Research Institute, Fujian, China *Email: pyao@mail.tsinghua.edu.cn

***Abstract*—This paper proposes a region-of-interest (ROI)-aware ReRAM accelerator supporting the microscaling (MX) datatype for next-generation AR/VR. It efficiently combines real-time gaze tracking with regional visual processing through three key features: (1) an operating-in-ReRAM ROI architecture with unified dual-mode macros for efficient MAC and special-function (SF) acceleration; (2) a distribution-aware MX (DA-MX) flow that mitigates outlier-induced scale deviation and inter-block dynamic-range loss in high-dynamic-range (HDR) data; and (3) a spatiotemporal differential execution (SDE) pipeline that eliminates redundant computation across consecutive ROI frames. Fabricated in 28 nm CMOS, the chip integrates 73 Mb ReRAM and achieves 22.31 TFLOPS/W system energy efficiency at MXINT8, with up to 88.5× FoM improvement over prior designs.**

***Index Terms*—computing-in-memory, ReRAM, edge accelerator, AR/VR, region-of-interest**

I. Introduction

Next-generation AR/VR demands computation-intensive ML-based image enhancement [1], [2] and scene analysis [3], [4] under strict power constraints. Restricting large-scale high-dynamic-range (HDR) processing to the region-of-interest (ROI) via lightweight gaze tracking [5] significantly reduces computational complexity, yet conventional architectures still incur prohibitive data-movement overhead. Computing-in-memory (CIM) is a promising solution with high energy efficiency. However, ROI-aware AR/VR requires diverse operators and high-precision HDR computing across consecutive frames, raising new challenges for CIM (Fig. 1): (1) Beyond MAC operators, ROI-aware AR/VR pipelines require special functions (SFs) such as spatial transformation [5] and feature normalization [1]. To achieve acceptable accuracy, SF execution logic incurs significant area and power overhead in current designs [6]. (2) MXINT [7] enables hardware-efficient HDR MAC with narrow bit-width, but precision degradation remains a challenge in AR/VR HDR processing. Sparse high-luminance outliers cause scale deviation in MX blocks, and inter-block accumulation with mismatched scales further compresses the effective dynamic range. (3) Consecutive ROI frames exhibit strong similarity across both temporal and spatial dimensions. Frame-to-frame gaze stability causes temporally redundant inference, while ROI boundary shifts leave the overlapped region redundantly reprocessed, incurring significant latency and energy overhead.

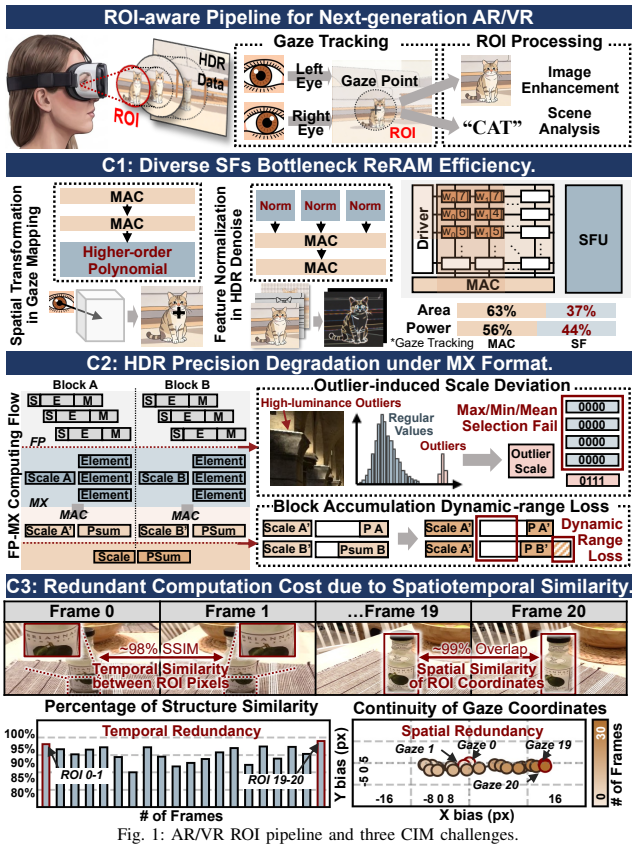


Fig. 1: AR/VR ROI pipeline and three CIM challenges.

II. Proposed Design

Among CIM technologies, ReRAM stands out for its non-volatility, high density, and rapid wakeup response, enabling all weights to be stored on-chip and completely eliminating off-chip DDR memory. In this work, we present a ReRAM-based ROI-aware edge accelerator with three key features to address the above challenges: (1) an operating-in-ReRAM ROI architecture that adopts unified dual-mode macros and a slope-aware iterative indexing scheme to efficiently accelerate both MAC and SF operators with adaptive precision; (2) a distribution-aware MX (DA-MX) processing flow that suppresses outlier-induced scale deviation via the dominant exponent selection and eliminates inter-block accumulation dynamic-range loss via the range-adaptation adder tree; (3)

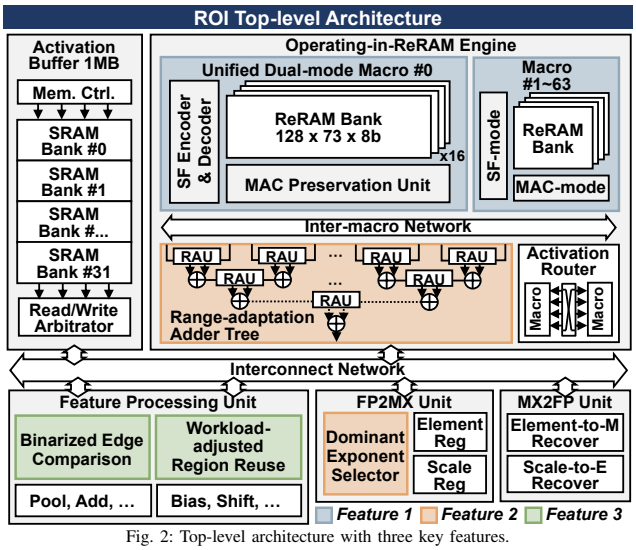


Fig. 2: Top-level architecture with three key features.

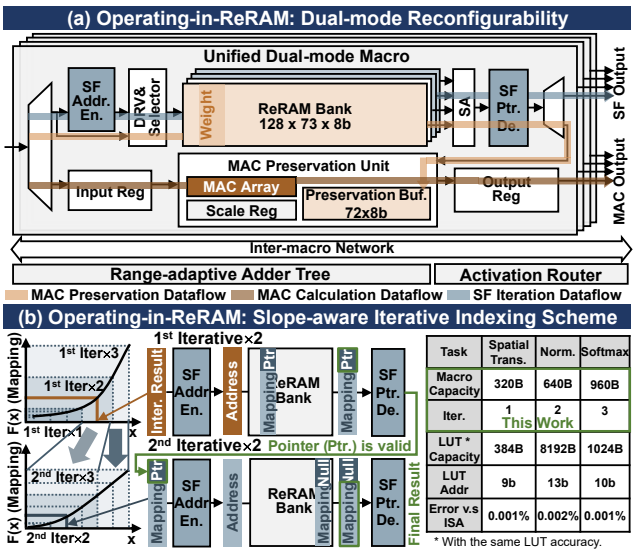


Fig. 3: The operating-in-ReRAM ROI architecture implements unified dual-mode macros for both MACs and SFs, and the slope-aware iterative indexing for adaptive SF precision.

a spatiotemporal differential execution (SDE) pipeline that reduces redundancy by comparing binarized edge features and reusing regions adjusted by workloads. Fig. 2 illustrates the top architecture: the chip comprises an operating-in-ReRAM engine with 64 dual-mode macros, FP-MX units for DA-MX, a feature processing unit for SDE, and a 1 MB activation buffer. The total effective weight storage is 73 Mb, with all SF results and MX block scales stored on-chip.

A. *Operating-in-ReRAM ROI Architecture*

Beyond standard activation functions such as GeLU and softmax, ROI-aware AR/VR pipelines introduce more SF operators from sensor and frame processing. For example, gaze tracking [5] requires spatial transformation to convert the gaze point from camera coordinates to screen coordinates, and HDR denoising [1] requires normalization of multi-modal data including texture, depth, and color. Conventional dedicated SFU logic incurs significant overhead. Existing LUT-based approaches [6] use fixed segmented mappings that require per-function customization of segmentations to cover the diverse slopes of different SFs, resulting in varying entry counts and address decoder widths that incur large area overhead. A unified, CIM-friendly, and high-precision solution is required.

We propose the operating-in-ReRAM ROI architecture with reconfigurable dual-mode macros and a slope-aware iterative indexing scheme. Fig. 3a details the macro structure and the dataflow. Each macro contains 16 banks of 128×73×8b 1-bit ReRAM cells and a 72×1b×8b MAC unit. Beyond the essential MAC computation modules, the macro only adds a lightweight SF encoder/decoder. The two modes run in parallel via decoupled dataflows.

Fig. 3b illustrates the slope-aware iterative indexing, which performs coarse-to-fine indexing through multiple iterations. Each iteration accesses a fixed number of entries with a progressively narrower range and finer resolution. A 2-bit pointer embedded in each entry determines at runtime whether the current slope requires further iterations, enabling a unified indexing process that adapts to different SFs without changing the input address width. In high-slope regions, the pointer directs the next iteration to finer-resolution entries, while in flat regions a null pointer terminates the indexing in a single pass. Each iteration level contains 256 entries, occupying 256×10b of bank capacity.

The evaluation results on three representative SFs are summarized in the table of Fig. 3b, including the third-order polynomial gaze spatial transformation $\mathbf{p}_s = \sum_{i+j\leq 3} a_{ij} x^i y^j$, the normalization function $1/\sqrt{x}$, and the softmax exponential e^x. For $1/\sqrt{x}$, 2 iterations with 256 fine entries achieve 0.002% maximum error. A uniform LUT achieving the same accuracy requires ≥8192 entries with a 13-bit address, consuming 10.7× storage and necessitating a wider address decoder. Instead of the conventional LUT, the proposed scheme keeps the input address at 8 bits for all SFs. The SF mode adds only 5% area and 6% power to the total chip overhead.

B. *Distribution-Aware MX (DA-MX) Processing Flow*

In AR/VR HDR processing, the floating-point (FP) format can represent luminance spanning several orders of magnitude. To enable hardware-efficient CIM, FP activations are converted to MXINT for MAC. However, two precision challenges arise. First, within each MX block of k elements, sparse high-luminance outliers from light sources and specular highlights dominate the E8M0 shared scale, degrading the precision of the activation distribution. Second, accumulating partial sums across blocks with different scales compresses the effective dynamic range of smaller-scale operands. Prior input-domain alignment [8], [9] compounds the error through narrower bit-width truncation, while two-level FP accumulation [10] introduces additional FP adders and control overhead.

DA-MX addresses both problems through an end-to-end FP-MXINT processing flow with the dominant exponent selection for input scale determination and the range-adaptation adder tree for partial-sum accumulation, as detailed in Fig. 4. Input

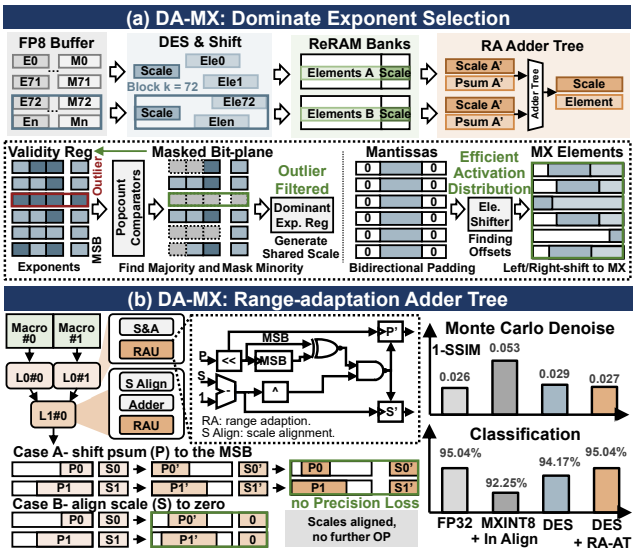


Fig. 4: The DA-MX applys dominant exponent selection with intra-bank mantissa alignment and range-adaptation adder tree for inter-bank accumulation.

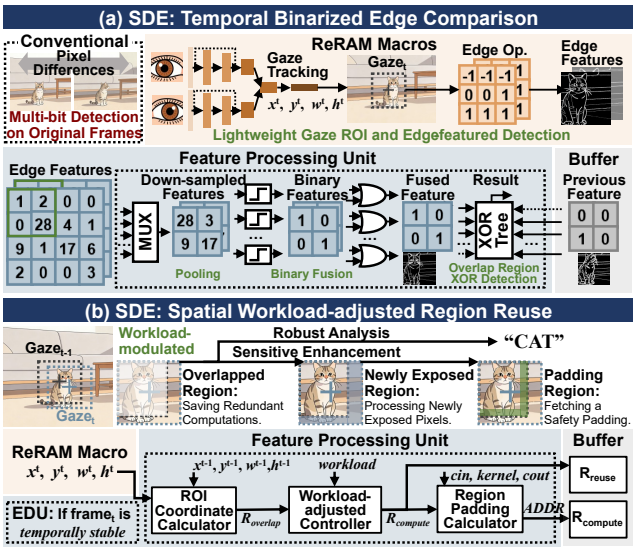


Fig. 5: The SDE pipeline leverages temporal binarized edge comparison and spatial workload-adjusted region reuse to filter out redundant computations.

activations are stored as E4M3 FP8 in the activation buffer and converted to MXINT8 for computation. To maximize ReRAM computation efficiency, block size k is set to the bank parallelism, which is 72 in this design. Each bank has 73 columns, where 72 elements share one E8M0 scale stored in the remaining column.

Fig. 4a shows the dominant exponent selection flow. It processes the exponents of the 72 inputs one bit-plane at a time. In each cycle, the 72 bits at the current bit position are collected into a validity register. A popcount comparator determines whether 0 or 1 is more frequent among the valid bits and clears the less frequent value. Only the dominant bits are processed in subsequent cycles. After 4 cycles, the remaining valid bits share the dominant exponent as the scale. Sparse luminance outliers are isolated in the cleared bits. In the subsequent shift phase, each mantissa is bidirectionally zero-padded to 8 bits with the sign bit preserved at the MSB. It is then left or right-shifted by its exponent offset from the shared scale to produce the MXINT8 representation. DA-MX centers the dominant data distribution within the effective bit-width. The conversion is fully pipelined with the bit-serial MXINT8 computation of the previous block. To support wider formats, the shift phase can be partially parallelized with the comparison phase to hide additional latency.

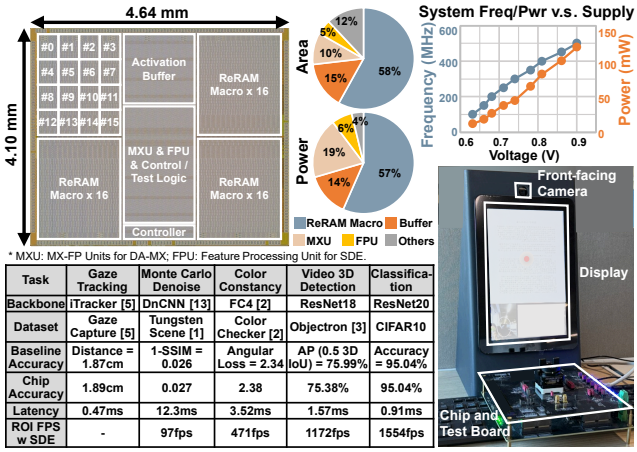
The range-adaptation adder tree in Fig. 4b mitigates dynamic-range loss during inter-block accumulation. At each tree level, a range-adaptation module left-shifts each partial sum until its MSB changes or the scale reaches zero, maximizing representable headroom. A subsequent scale-alignment module then aligns the different local scales before addition, preventing smaller-scale operands from being truncated when merged with larger-scale results. The accuracy comparison with ablation is summarized in Fig. 4b. The full DA-MX scheme incurs only a 0.001 1-SSIM degradation on Monte Carlo denoising compared with FP32 software, and maintains 95.04% baseline accuracy on classification.

C. Spatiotemporal Differential Execution (SDE) Pipeline

Consecutive ROI frames contain significant redundancy. Temporally, when the scene is static or gaze moves slowly, successive frames carry nearly identical content. Spatially, when the ROI boundary shifts, the overlapping region is redundantly reprocessed. Prior methods operate at the full-frame level with workload-uniform policies, introducing non-negligible area and computation overhead [11], [12].

SDE suits ROI-aware processing by detecting redundancy with binarized edge comparison and reusing overlapped regions adjusted by workloads (Fig. 5). For temporal redundancy, the binarized edge comparison is pipelined with gaze tracking and reuses the MAC and pooling operators of the ROI architecture, as shown in Fig. 5a. Prewitt operators, executed as standard convolutions on the macro, extract edge features from consecutive ROI frames. Pooling-based downsampling then reduces minor fluctuations and controls the detection sensitivity. The downsampled edge features are then binarized via threshold comparison for efficient bitwise evaluation. Both the pooling size and threshold are determined via parameter search. For analysis tasks, the pooling size is set to 2 and the threshold to 63. For enhancement tasks, the pooling size is set to 1 and the threshold to 31. A bitwise XOR tree compares the current and previous maps to produce a scalar frame-divergence metric in $O(\log N)$. Frames with similarity exceeding 90% are forwarded to the spatial detection stage.

The spatial region reuse illustrated in Fig. 5b depends on the workload characteristic and the overlap ratio between consecutive ROI frames. For analysis tasks, where the newly exposed region does not affect the global prediction, the overlapped region is directly skipped when the ratio exceeds 60%. For enhancement tasks, where every output pixel contributes to image quality, the ratio is raised to 80% and a configurable padding region is applied around the newly exposed region to guarantee lossless processing at the boundary. The padding



* MXU: MX-FP Units for DA-MX; FPU: Feature Processing Unit for SDE.

Task	Gaze Tracking	Monte Carlo Denoise	Color Constancy	Video 3D Detection	Classification
Backbone	iTracker [5]	DnCNN [13]	FC4 [2]	ResNet18	ResNet20
Dataset	Gaze Capture [5]	Tungsten Scene [1]	Color Checker [2]	Objectron [3]	CIFAR10
Baseline Accuracy	Distance = 1.87cm	1-SSIM = 0.026	Angular Loss = 2.34	AP (0.5 3D IoU) = 75.99%	Accuracy = 95.04%
Chip Accuracy	1.89cm	0.027	2.38	75.38%	95.04%
Latency	0.47ms	12.3ms	3.52ms	1.57ms	0.91ms
ROI FPS w SDE	-	97fps	471fps	1172fps	1554fps

Fig. 6: Die photo with area/power breakdown, system frequency/power vs. supply voltage, demo system, and per-task measurement results.

	This Work	ISSCC'26 [8]	ISSCC'25 [14]	VLSI'25 [15]	JSSC'24 [16]	ISSCC'24 [17]
Technology	28nm	22nm	16nm	16nm	16nm	22nm
NVM Device	ReRAM	-	-	-	MRAM	ReRAM
Chip Hierarchy	System	System	System	System	System	Macro
Compute Mode	MAC/SF Dual-mode	Digital+CIM+ TCAM	Digital	Digital	Digital	CIM
Application	AR/VR Gaze Tracking + Enhancement/ Analysis	Automatic Speech Recognition	Video Detection	Image Processing	XR Analysis	Edge AI
Weight Capacity (Mb)	73	1.93	16.125	10.32	32	16
Supply Voltage (V)	0.63-0.9	0.55-1.0	0.66-1.09	0.63-1.08	0.65-0.8	0.7-0.8
Frequency (MHz)	100-500	46.4-582	100-480	100-400	360	-
System-level Energy Efficiency (TOPS/W, TFLOPS/W)	22.31 (MXINT8)*1 26.92 (INT8) 40.63 (INT4)	10.24 (MXFP8)	8.9 (INT8)	10.0 (INT8)	2.68 (INT8)	-
Macro-level Energy Efficiency (TOPS/W, TFLOPS/W)	31.45 (MXINT8)*1 37.57 (INT8) 52.90(INT4)	16.72 (MXFP8)	-	-	-	28.7 (FP16)
FoM*2	1.22E4	2.39E2	5.57E2	5.34E2	1.38E2	5.38E3

*1: Measured @ 0.63V, 100MHz, 25% input sparsity. *2: FoM = TOPSN/W x Mb/mm²; TOPSN/W = IN bits x W bits x TOPS/W. Calculated at the system level for SoC. The area is scaled to 28nm technology.

Fig. 7: Performance comparison with prior designs.

width is set to 18 pixels for Monte Carlo denoising. SDE results are analyzed in Section III.

III. Measurement Results

Fig. 6 shows the die photo, area/power breakdown, performance scaling, demo system, and per-task results. The chip occupies $4.64\times4.10\,\text{mm}^2$ in 28 nm CMOS and integrates 92.45 Mb ReRAM (73 Mb effective weights, remainder for ECC and dummy cells). It operates at 100–500 MHz with a 0.63–0.9 V supply. The demo runs the full gaze-tracking [5] and AR/VR pipeline on-chip without off-chip DDR. Four representative AR/VR tasks are evaluated with per-task accuracy and latency listed in Fig. 6, including Monte Carlo denoising [1], [13], color constancy [2], video 3D detection [3], and classification [4]. The chip delivers 97 fps on denoising and up to 1554 fps on classification with the SDE. SDE provides 32.7% average computation reduction across the four tasks.

Fig. 7 compares this work with recent CIM and digital accelerators [8], [14]–[17]. The chip achieves 22.31 TFLOPS/W full-system energy efficiency at MXINT8. As the design stores all weights on-chip without DDR, we adopt a storage-density-aware $\text{FoM} = \text{TOPS}_\text{N}/\text{W} \times \text{Mb/mm}^2$ that jointly captures energy efficiency and weight density. This work achieves the highest system-level energy efficiency and a SOTA FoM improvement ranging from 2.3× over prior CIM to 88.5× over conventional digital architectures.

IV. Conclusion

This paper presented a 28 nm 73 Mb ReRAM-based ROI-aware edge accelerator achieving 22.31 TFLOPS/W system energy efficiency. The dual-mode macro supports diverse SFs with only 5% area and 6% power overhead. The DA-MX preserves FP32-level accuracy with only 10% area overhead. The SDE pipeline eliminates 32.7% of redundant computation across consecutive ROI frames. Silicon measurements on four AR/VR tasks demonstrate SOTA energy efficiency and up to 88.5× FoM improvement over prior designs.

Acknowledgment

This work was supported by the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2021ZD0201200), the NSFC (62422405 and 62495103), the Shanghai Municipal Science and Technology Major Project, and the Beijing Advanced Innovation Center for Integrated Circuits.

References

[1] S. Bako *et al.*, "Kernel-predicting convolutional networks for denoising Monte Carlo renderings," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.
[2] Y. Hu *et al.*, "FC4: Fully convolutional color constancy with confidence-weighted pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 330–339.
[3] A. Ahmadyan *et al.*, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7822–7831.
[4] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[5] K. Krafka *et al.*, "Eye tracking for everyone," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.
[6] W. Xie *et al.*, "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," in *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*, 2025, pp. 125–128.
[7] B. D. Rouhani *et al.*, "Microscaling data formats for deep learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
[8] W. Ren, M. Li *et al.*, "A 22nm 1.87ms/frame streaming multi-speaker ASR accelerator leveraging contextual-aware redundancy skipping with 2D-writable microscaling compute-in-memory and similarity-aware TCAM design," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2026.
[9] W.-S. Khwa *et al.*, "A 16nm 216kb, 188.4TOPS/W and 133.5TFLOPS/W microscaling multi-mode gain-cell CIM macro edge-AI devices," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, 2025, pp. 1–3.
[10] Z. Yue *et al.*, "A 28nm 47.3TFLOPs/W 894mJ/Inference visual autoregressive accelerator with differential-amplifier speculation and chain-reaction-like parallel generation," in *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, 2026, pp. 316–318.
[11] M. Wang *et al.*, "VideoTime^3: A 40-μJ/frame 38 FPS video understanding accelerator with real-time DiffFrame temporal redundancy reduction and temporal modeling," *IEEE Solid-State Circuits Letters*, 2023.
[12] X. Wang *et al.*, "InterArch: Video transformer acceleration via inter-feature deduplication with cube-based dataflow," in *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC)*, 2024.
[13] K. Zhang *et al.*, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
[14] Y.-C. Ding *et al.*, "A 16nm 5.7TOPS CNN processor supporting bi-directional FPN for small-object detection on high-resolution videos," in *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, 2025, pp. 1–3.
[15] Y.-T. Chen *et al.*, "CATTUS: A 4K-UHD 30FPS deep image processor for channel attention equipped U-Net acceleration in 16nm FinFET," in *IEEE Symposium on VLSI Technology and Circuits*, 2025, pp. 1–3.
[16] A. S. Prasad *et al.*, "Siracusa: A 16nm heterogeneous RISC-V SoC for extended reality with at-MRAM neural engine," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 7, pp. 2055–2069, 2024.
[17] T.-H. Wen *et al.*, "A 22nm 16Mb floating-point ReRAM compute-in-memory macro with 31.2TFLOPS/W for AI edge devices," in *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, 2024, pp. 580–582.