H3DFACT: Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations

***Abstract*—Disentangling attributes of sensory signals is central to sensory perception and reasoning and hence is a critical task for future cognitive and neuro-symbolic AI systems. An elegant approach to represent this intricate factorization is via high-dimensional holographic vectors in the context of brain-inspired vector symbolic architectures. However, holographic factorization involves iterative computation with high-dimensional matrix-vector multiplications and suffers from non-convergence problems.**

In this paper, we present H3DFACT, the first heterogeneous 3D integrated in-memory compute engine capable of efficiently factorizing high-dimensional holographic representations. H3DFACT exploits the computation-in-superposition capability of holographic vectors and the intrinsic stochasticity associated with memristive-based 3D compute-in-memory. Evaluated on large-scale factorization and holographic perceptual problems, H3DFACT demonstrates superior capability while solving at least five orders of magnitude larger problems, as well as substantially lowering the computational time and space complexity. H3DFACT achieves 5.5× compute density, 1.2× energy efficiency and 5.9× less silicon footprint compared to iso-capacity 2D designs.

I. INTRODUCTION

The brain's remarkable ability to reason and make sense of the world relies heavily on its capacity to disentangle sensory attributes. This intricate process involves the factorization of various sensory inputs, such as vision, hearing, touch, and more, into distinct perceptual features [1]. This factorization not only aids in perception but also serves as the foundation for higher-order cognitive functions like problem-solving, decision-making, and abstract thinking, thus serving as a crucial component for neuro-symbolic AI [2].

An elegant mathematical approach to represent this intricate factorization is via high-dimensional holographic vectors in the context of brain-inspired vector symbolic architecture (VSAs) [3]. Each sensory attribute or feature is encoded and processed using a unique holographic vector, thereby creating distinct and separable representations. These vector representations can be manipulated using a set of algebraic operations. For example, an object with multiple attributes can be described by the element-wise multiplication of all vectors representing these attributes. The factorization problem in turn is concerned with decomposing a product vector into its constituent attribute vectors. This is a hard combinatorial search problem that arises when dealing with products of complex attribute structures [4].

The compositional nature of holographic vector representations gave rise to an efficient factorization algorithm, *resonator network* [5], that equip with superior ability to bridge cognitive gaps in neuro-symbolic AI, by accepting holographic perceptual representations from a neural network and factorizing them for symbolic reasoning [6]. Resonator network is able to perform search in superposition, which allows for simultaneous exploration of a product's constituent elements. This factorization procedure exhibits characteristics akin to dynamic systems, engaging in an iterative computation flow with high-dimensional matrix-vector multiplications (MVMs). It also relies on stochastic exploration strategies, which aim to circumvent the potential pitfalls of falling into limit cycle problems. These key features make factorization amenable to computing platforms that enable compute-in-memory (CIM) and are inherently stochastic, such as memristive devices.

Recently, an in-memory factorizer using the resonator network was proposed [7], where each individual die contains a 2D CIM array to accelerate a specific MVM operation. This approach, however, does not exploit the full potential of CIM; it incurs considerable cost due to the increased silicon area and substantial data communication between different dies in each iteration. Our goal is to achieve highly efficient holographic factorization by capitalizing on the capabilities offered by emerging memory technologies, which permit integration of multiple heterogeneous arrays in 3D-stacked configuration.

In this paper, we propose H3DFACT, the first heterogeneous 3D (H3D) integrated CIM factorizer for high-dimensional holographic vector representations. H3DFACT features a hybrid memory design, which integrates analog RRAM computation with digital-SRAM components. The RRAM tier is used to efficiently process MVM operations and is designed using a legacy technology node to support relatively high programming voltages. The RRAM's peripheral circuitry, on the other hand, is placed on a separate tier and is integrated with SRAM units using a more advanced node. The integration of these tiers via an H3D configuration leads to major improvements in silicon area and power consumption. Furthermore, the non-deterministic nature of the described memory elements enhances factorization convergence and operational capacity. When compared to iso-capacity 2D designs, H3DFACT demonstrates superior efficiency in terms of power, performance, and area.

This paper, therefore, makes the following contributions:

- We propose the first H3D integrated CIM accelerator, H3DFACT, for efficient and scalable factorization of high-dimensional holographic representations.
- We present a hybrid-memory design that combines the merits of RRAM computation in legacy nodes (40 nm) and digital-SRAM components in advanced nodes (16 nm).
- We demonstrate that H3DFACT improves factorization accuracy and operational capacity by up to five orders of magnitude due to inherent stochasticity, with 5.5× compute density, 1.2× energy efficiency, and 5.9× less silicon footprint compared to iso-capacity 2D designs.

II. Background and Motivation

This section presents high-dimensional vector operations for perceptual encoding and factorization, and motivates the proposed 3D integrated CIM solution designed for factorization.

A. High-Dimensional Holographic Vector Operations

In high-dimensional holographic vector operations, atomic features and patterns can be encoded using randomly generated vectors (*item vectors*) $x_i \in \{-1, +1\}^D$, where D can be in the range of thousands. Due to the randomness and holographic nature of high-dimensional vectors, item vectors are therefore quasi-orthogonal, i.e., dissimilar, allowing for the disambiguation of the different represented features. These vectors can be manipulated using the following operations [8]: (1) element-wise multiplication ($\odot$), which can be used for "binding" item vectors to create a product and also for "unbinding" a product to retrieve item vectors; (2) element-wise addition ($[+]$), which computes the superposition of multiple products; (3) permutation (ρ), which changes the ordering of vector elements to capture the sequence of the feature.

B. Factorization & Resonator Network

We show here how holographic vectors are used to encode the compositional structure of objects and how the resonator network works to decode the contents of this structure via factorization. Consider an example of encoding visual objects, which are characterized by four attributes ($F = 4$): shape, color, vertical position, and horizontal position. As illustrated in Fig. 1a, each of these four attributes corresponds to a different M-sized codebook of randomly generated item vectors. This way, an object vector can be formed through the binding of vectors from these codebooks.

The resonator network (factorization) works in the opposite direction. That is, it seeks to decompose an object vector into its constituent attribute vectors. The only inputs given to this algorithm are the composed object vector along with the individual codebooks of features. The algorithm compositionally searches through these codebooks to find the exact feature vectors. The following state-space equations describe this search (Fig. 1b):

$$\hat{x}(t+1) = g\big(XX^\top(s \odot \hat{c}(t) \odot \hat{v}(t) \odot \hat{h}(t)\big); \quad X = [x_{cir}\ x_{tri}\ \ldots]$$

$$\hat{c}(t+1) = g\big(CC^\top(s \odot \hat{x}(t) \odot \hat{v}(t) \odot \hat{h}(t)\big); \quad C = [c_{blue}\ c_{red}\ \ldots]$$

$$\hat{v}(t+1) = g\big(VV^\top(s \odot \hat{c}(t) \odot \hat{x}(t) \odot \hat{h}(t)\big); \quad V = [v_{top}\ v_{bottom}]$$

$$\hat{h}(t+1) = g\big(HH^\top(s \odot \hat{c}(t) \odot \hat{v}(t) \odot \hat{x}(t)\big); \quad H = [h_{left}\ h_{right}]$$

Here, t is a time step; s is the object vector; $\hat{x}$, $\hat{c}$, $\hat{v}$, and $\hat{h}$ hold the predicted values of the features x, c, v, and h, respectively.

We observe that MVM operations dominate most of the computation time in the factorization algorithm. As shown in Fig. 1c, MVM operations within similarity and projection steps account for 80% of the total computation time. This result establishes a clear motivation for adopting a CIM design approach, which provides ways for MVM operations to always execute in a constant time irrespective of the problem size.

Another motivation for using the CIM design approach is to address a major issue with the scaling of the factorization accuracy. Specifically, we observe a significant drop in the factorization accuracy with increasing the problem size (Fig. 1c). This accuracy drop is due to the limit cycle problem, which can be a limiting factor for large-scale factorization. One effective solution is to introduce stochasticity, which helps to break free of limit cycles and thus explore a substantially larger solution space. CIM devices are inherently stochastic; therefore, they provide a natural way for implementing this solution.

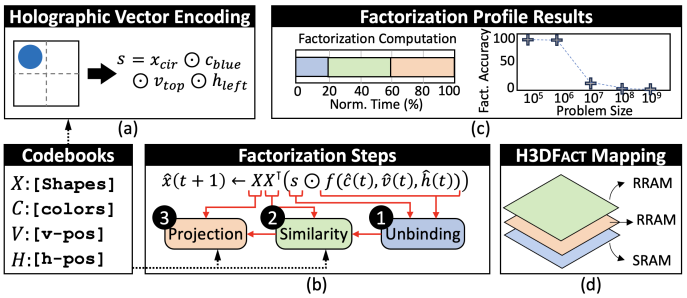


Fig. 1: Computational Primitives of the Holographic Vector Encoding and Factorization. (a) Vector encoding of a visual object. (b) Algorithmic flow of the factorization. (c) Results of profiling the factorization operations. (d) A schematic of the proposed H3D integrated factorizer.

C. Heterogeneous 3D CIM Acceleration

Prior 3D integrated hardware designs have mainly focused on accelerating CNNs [9] or Monolithic 3D integration [10]. In contrast, H3DFact tackles a different MVM workload, heavily used in high-dimensional cognitive systems, and maps different components of this workload to hybrid RRAM/SRAM memory tiers (Fig. 1d). Moreover, H3DFact provides flexibility in designing with hybrid technology nodes, thus leading to significant improvements in the compute density, energy efficiency, and silicon footprint compared to iso-capacity 2D designs.

III. Compute-In-Memory Primitives

This section first presents a detailed circuit-level view of H3DFact memory tiers, and then discusses the benefits of H3DFact inherent stochasticity to factorization convergence.

A. RRAM Tier

Fig. 2a provides a macro-level overview of the RRAM tier, depicting multiple arrays on a single tier. Each array is equipped with circuitry capable of executing MVM in the high-dimensional bipolar space ($\{-1, +1\}^D$). This circuitry includes a specialized -1's counter and an adder that processes bipolar quantities [11]. Note that existing array designs for VSAs often fall short in fully supporting the bipolar space, as they frequently map a bipolar element $\{-1, +1\}$ to a single-bit quantity [12]. This approach, however, is not suitable for the factorization algorithm, which seeks to accumulate both positive and negative quantities within its computational flow.

The operation of the RRAM involves setting and resetting using high-voltage signals, necessitating the inclusion of isolation switches to protect peripherals against these high voltages. Voltage regulation is achieved through a PMOS device connected to a power supply (AVDD) along with an operational amplifier (Fig. 2b). $VTGT$ represents the target sensing voltage in the

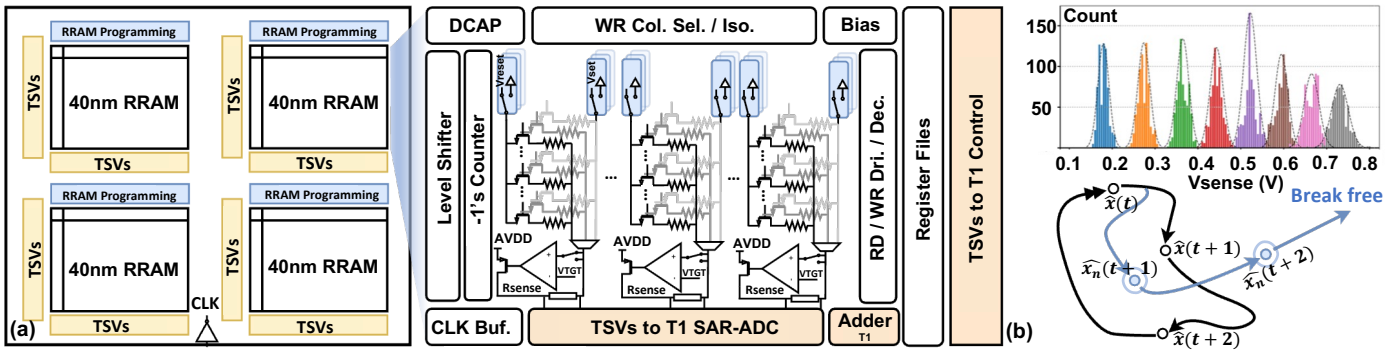


Fig. 2: H3DFACT Array-Level Components. (a) Legacy node RRAM tier-level view and building blocks for a single RRAM array. (b) The inherent stochasticity of H3DFACT helps break limit cycles and benefit factorization convergence.

sensing path. Additionally, a current-sensing resistor (Rsense) is incorporated, enhancing PVT (Process-Voltage-Temperature) immunity. Given that RRAM can be subject to frequent power-switching events, the design allows for different power-off modes, including a full shutdown, while enabling other tiers to remain active. These functions were experimentally validated using a chip fabricated with a 40 nm technology node [13].

B. Digital-SRAM Tier

The interaction between the RRAM and the peripherals takes place through digital circuitry that includes an analog-to-digital (ADC) converter, an adder, and a controller (depicted by the orange-colored blocks in Fig. 2a). One of the advantages of heterogeneous integration is the integration of systems with different technology nodes [14]. An area mismatch between RRAM and its peripherals has resulted in MUX-sharing of the RRAM sensing [13]. To fully unleash the system performance, digital components in H3DFACT are thus designed in 16nm advanced node, enabling a sensing path for each RRAM's output.

We adopt a hybrid-computing scheme for the frequently updated operation in the factorization as the write operation for NVM is notorious for its humongous overhead [15]. The hybrid-computing scheme utilizes XNOR logic gates for bit-wise unbinding operation [16]. This is driven by constant memory write operation in unbinding updates for different time steps in factorization. In addition to the hybrid-computing scheme, a hybrid-memory (SRAM near-memory) scheme for buffering TSVs data transfer in the H3D design, which will be further explained in Sec. IV. To reduce the overhead of Through-silicon vias (TSVs), we only enable connections across different tiers in their input and output ports. For instance, connections only exist at the input row and output column for each RRAM array. This approach follows recent H3D design as excessive TSVs can severely damage not only the system-level PPA, but also RRAM ON/OFF ratio [17]. However, this approach will require only one RRAM tier being activated. In Sec. IV, we further discuss the architectural impact of RRAM activation.

C. Stochastic Factorizer

The unsupervised nature of the factorization's deterministic search could result in checking the same sequence of solutions multiple times across iterations, preventing convergence to the optimal solution in limited cycles. One of the key insights from H3DFACT is that the intrinsic stochasticity associated with memristive devices can substantially reduce the occurrence of such limit cycles. As shown in Fig. 2b, in-memory MVM readout results in a stochastic similarity vector with all the PVT variations aggregated altogether. The $\hat{x_n}(t+1)$ indicates noisy $\hat{x}$ at the $t+1$ time step. The hardware stochasticity enables the factorizer to break free of potentially being stuck at limited cycles and thus has the ability to explore a substantially larger space, demonstrating the potential to leverage device-level dynamics as a valuable source for application performance.

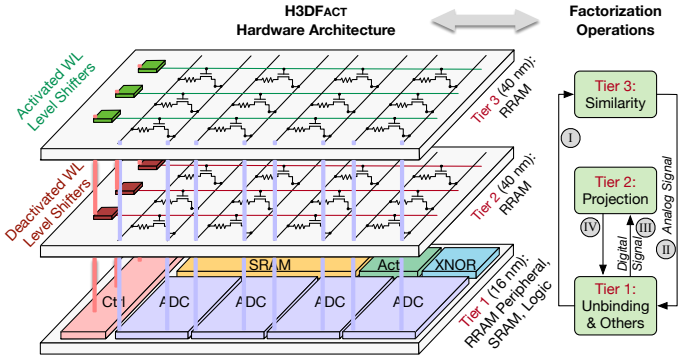


Fig. 3: H3DFACT Architecture and Control Scheme. The factorization computation kernels are partitioned among three vertical tiers. The control scheme for activating only one tier of RRAM CIM arrays when all RRAM tiers share the same vertical interconnects. Turning off the power to WL level shifters (red) will deactivate the current flow in the corresponding RRAM arrays.

IV. H3DFACT ARCHITECTURE

This section presents the H3DFACT architecture, including the data flow, hardware design, interconnects, and floor plan.

A. Proposed H3DFACT Architecture

Factorization Workload Mapping. H3DFACT realizes factorization by partitioning its computational kernels into three tiers, in which similarity calculation, projection, digital operations, lie in tier-3, tier-2, and tier-1, respectively (Fig. 3). This design choice is related to the fact that the data is traversing in a digital or analog manner, where step I is the unbinding results for similarity calculation, step II is the similarity outputs that are represented with analog current, step III is the 4-bit digital

TABLE I: H3DFACT Interconnect Specifications.

TSV Diameter	TSV Pitch	TSV Oxide Thickness	TSV Height	Hybrid Bonding Pitch	Hybrid Bonding Thickness
2 μm	4 μm	100 nm	10 μm	10 μm	3 μm

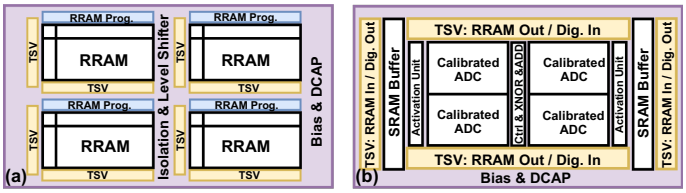


Fig. 4: H3DFACT Floor Plan. (a) RRAM tier-2/3. (b) Digital tier-1.

result obtained from the similarity calculation, and step IV is the 1-bit digital data from projection. Analog data transfer from tier-3 to tier-1, and tier-2 to tier-1 is deemed to be negligible, as analog current can flow through the TSV one-shot. On the other hand, the 4-bit digital similarity results obtained from tier-1 is passed to projection tier-2 to avoid multibit digital value transmission degrading system performance Thus, H3DFACT is designed with similarity at the top, projection at the middle, and digital circuits with advanced node at the bottom.

Tier-2 & Tier-3 RRAM CIM. One set of RRAM peripherals is shared among both RRAM tiers (tier-2 and tier-3), such that the tier-to-tier interconnects from tier-1 are connected to all tiers. The implication is that only one RRAM tier can be active at once in this design, which explains the reasoning behind placing WL level shifters in each RRAM tier to control the activation of the RRAM tiers. Fig. 3 describes the control scheme of two tiers of RRAM. The RRAM WLs, BLs, and SLs between the tiers are effectively shorted in the vertical direction due to the sharing of peripherals. In order to activate only one tier of RRAM at once, as mentioned, the designed chip is capable of being totally shut down. This truthfully turns off the transistors in the 1T1R cells, so the RRAM cells in idle tiers do not contribute any column current.

Tier-1 SRAM Digital Compute. We adopt SRAM in tier-1 to support greater-than-one factorization batch size. Considering a batch size of 100, after similarity calculation (tier-3), the similarity outputs propagate to tier-1 for analog-digital conversion. If without SRAM buffering, the tier-1 ADC-output signals are sent to tier-2 for projection calculation, which will violate single-RRAM tier activation because tier-3 is still computing similarity for data in the same batch. Therefore, we propose to adopt SRAM in tier-1 to serve as buffers to support large batch factorization computation.

Design Methodology Generalization. The H3DFACT architecture can accommodate the varying parameters of resonator networks. Since resonator networks are parametrized with high dimensional vector dimensions M and F, the H3DFACT is likewise parameterized with hardware dimensions m and f. In H3DFACT, the parameter m determines the number of rows of an RRAM-CIM array, and is set to 256. The parameter f determines the number of RRAM subarrays of each tier and is set to 4. After accommodating the vector size, multiple inputs may also be processed in parallel by different subarrays.

B. Tier-to-Tier Interconnects in H3DFACT

We summarize the parameters of the tier-to-tier interconnects for the H3DFACT design in Tab. I. These interconnect assumptions align with recent commercial designs such as AMD's 3D V-Cache [18]. For an RRAM array with M rows and N columns, the required number of TSVs for connecting to the RRAM peripheral tier (Tier 1) is M for WLs + N for BLs + $N/2$ for SLs. It is important to note that the choice of array size affects the TSV overhead, as larger arrays require fewer TSVs but tend to be more underused than smaller arrays. We choose to store the similarity matrix at each iteration in the same array to amortize the usage of TSVs.

Motivated by the nontrivial area overhead of TSVs, we choose analog CIM and SAR ADC in H3DFACT design to reduce the TSV area overhead. In analog CIM, the partial sum of a MVM operation is represented in analog current, thus transmitting the analog signal to an ADC located in tier-1 only requires one set of interconnects. Although RRAM-based digital CIM has also been proposed [19], one TSV would be required for every RRAM cell if the RRAM array and the digital adder trees were partitioned into different tiers, leading to an impractical 3D design. The choice of ADCs also impacts TSV overhead. A critical concern for using Flash ADC in 3D design is its parallel output. An N-bit Flash ADC requires N interconnects for tier-to-tier transmission; however, serializing the Flash ADC output would defeat its speed merit. Therefore, the output-serial nature of the SAR-ADC presents a better strategy to reduce TSV overhead in H3DFACT.

C. Floor Planning and Bonding of H3DFACT

We approximate the floor plan of each tier in the H3DFACT to check that the tiers along the 3D stack are indeed area-balanced and also to use as inputs to the thermal analysis (Sec. V-C). Areas of CIM arrays and their peripherals are estimated by the calibrated NeuroSim framework [20], whose estimations have been validated against RRAM-based CIM chips [13]. Areas of other digital modules are obtained from TSMC standard cell library.

The two RRAM tiers have identical structures, so we show the floor plan of one RRAM tier in Fig. 4a. Each RRAM subarray has a dimension of 256 $\times$ 256, with four designed for each tier. We can perform RRAM CIM in any particular subarray(s) by activating their corresponding WLs and BLs.

Fig. 4b shows the floor plan for the RRAM peripheral and SRAM digital compute tier. The controller and buffer are also placed at tier-1 to avoid many connections to other SoCs or packages, as the external pins and C4 bumps are on the bottom tier [9]. H3DFACT provides the flexibility to design RRAM peripheral circuitries in more advanced nodes [17], [21], thus we choose to assign each RRAM column with a 4-bit ADC. To validate, we quantize the similarity calculation to 4-bit, and observe no factorization accuracy drop while having even faster convergence than 8-bit ADC design (Sec. V-D).

Regarding bonding techniques between tier-to-tier TSVs, we consider a mix of face-to-face (F2F) and face-to-back (F2B). In F2B integration, TSVs bond multiple tiers. As TSVs penetrate

TABLE II: Accuracy Evaluation. Factorization accuracy and operational capacity comparison under different problem sizes.

	Factorization Accuracy (%)				Number of Iterations*			
	F=3		F=4		F=3		F=4	
	Baseline	H3D	Baseline	H3D	Baseline	H3D	Baseline	H3D
M=16	99.4	99.3	99.2	99.2	4	5	31	33
M=32	99.3	99.3	99.1	99.2	13	15	234	140
M=64	99.1	99.3	89.9	99.2	43	39	Fail	1347
M=128	96.9	99.3	0	99.2	Fail	108	Fail	17529
M=256	10.8	99.2	0	99.2	Fail	443	Fail	269931
M=512	0.2	99.2	0	99.2	Fail	1685	Fail	2824079

* Number of iterations required to reach at least 99% accuracy under different problem sizes.

through the silicon, the memory placement or the TSV usage gets restricted. While F2F does not pose any place and route restriction, it is impossible to integrate all three tiers using F2F integration [9], and a mix of F2F and F2B tier-to-tier connections are required for three-tier H3DFACT design.

V. H3DFACT EVALUATION

This section evaluates H3DFACT on factorization and holographic systems. We demonstrate that H3DFACT achieves improved accuracy, operational capacity, and hardware efficiency, and illustrate its robustness with RRAM chip validation.

A. Accuracy and Operational Capacity

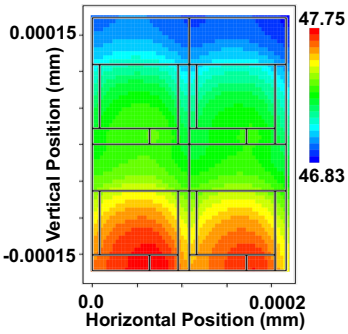
Accuracy Improvement. Tab. II compares the factorization accuracy of H3DFACT with baseline resonator network [5] under different number of attributes F and code vectors M. While both baseline network and H3DFACT achieve 99% accuracy under small M^F, H3DFACT substantially enhances and maintains 99% accuracy under high dimensionality, illustrating its improved scalability for larger factorization problem sizes.

Operational Capacity Improvement. Tab. II also shows the number of iterations required to solve a given problem size with accuracy of at least 99%. Compared with baseline resonator network [5], H3DFACT enables faster convergence and can solve problem sizes at least five orders of magnitude larger at 99% accuracy, illustrating H3DFACT capable of lowering computational complexity with improved operational capacity.

B. Hardware Efficiency

Monolithic 2D Baseline Design Setup. We analyze the merits of the proposed H3DFACT by comparing it with two 2D architectures: a hybrid RRAM/SRAM design and a fully SRAM design (Tab. III). For hybrid 2D design, all modules must be implemented in 40 nm to follow RRAM process as a monolithic 2D, and we assume that one set of RRAM peripherals is shared by two RRAM arrays with each RRAM column sensed by an ADC. For the fully SRAM design, all modules are scaled to 16 nm nodes. For a fair comparison, we assign the same amount of computing resources and parameters for all designs.

Silicon Footprint Reduction. Tab. III shows that fully SRAM design in 2D requires an area of $0.114mm^2$ with all components in 16 nm. The 2D RRAM/SRAM hybrid design occupies up to $0.544mm^2$ despite involving no TSV overheads due to limitation in current RRAM fabrication technology. In contrast, the advanced node scaling and vertical integration in H3DFACT allow a more compact footprint of $0.091mm^2$.



Attribute	Value
Number of tiers	3
PCB thickness	2 mm
Bumping thickness	100 μm
Package thickness	1 mm
TIM thickness	TIM1: 20 μm TIM2: 20 μm
Heat transfer coefficient	1000 (W/m²°C)
Ambient temperature	25°C

Fig. 5: Thermal Analysis. Thermal map of H3DFACT with its setup.

Even accounting for all three tiers, H3DFACT still provides appreciable reductions of 1.25× and 5.97× in total silicon cost compared to fully SRAM and hybrid 2D designs, respectively.

Compute Density and Energy Efficiency Improvement. Compared to 2D designs, H3DFACT operates at a slightly lower frequency (185 MHz) due to parasitic capacitance from TSVs and hybrid bindings, resulting in a slight throughput penalty as all designs have the same amount of compute resources. Nevertheless, as in Tab. III, H3DFACT still enjoys 1.2× higher compute density and 1.2× energy efficiency by scaling RRAM peripheral and digital components from 40 to 16 nm. Even when all modules are scaled to 16 nm in the 2D fully deterministic digital SRAM baseline, H3DFACT still attains comparable energy efficiency with 5.5× higher compute density and 3.5% higher factorization accuracy due to the associated intrinsic stochasticity (Fig. 2c).

Compare with Other Factorization Accelerators. Compared with recent PCM-based in-memory factorization [7], H3DFACT achieves 1.78× throughput and 1.48× energy efficiency under the same silicon area by virtual of 3D stacking and improved compute density, with >99% factorization accuracy.

C. Thermal Evaluation

Thermal Simulation Setup. One critical aspect in designing 3D IC is thermal profile, as stacking active dies can cause the maximum junction temperature to rise, leading to performance and reliability loss [22]. Our chip-level thermal setup includes hybrid bonds and TSVs to connect Tiers 1-3, C4 bumps to connect Tier 1 to package, and thermal interface material (TIM) at top for cooling. The parameters are summarized in Fig. 5.

Thermal Analysis. We use HotSpot [23] to evaluate the H3DFACT thermal where we assign the power density associated with each component according to the floorplans (Fig. 4). As in Fig. 5, the tier temperatures for H3DFACT range from 46.8 °C to 47.8 °C, where the 2D design is 44 °C. With cooling is more effective at center and high power density lies in the southern of each macro, as expected, there exist slight temperature increases toward the die southern region. Overall, 3D stacking in H3DFACT does not severely affect the reliability of RRAM (RRAM retention and drift will be seriously impacted in temperatures exceeding 100 °C [24]).

D. Robustness Evaluation and Chip Validation

Convergence Speedup. Lowering ADC precision can reduce hardware costs and enable faster convergence of holographic

TABLE III: Hardware Performance Evaluation. Hardware resource and performance comparison between 2D and H3DFACT Designs.

Design Choice	Hardware Resource							Performance					
	Technology (RRAM)	Technology (RRAM Peripheral)	Technology (Digital)	Unbinding Operation	Similarity & Projection Operation	ADC Count	TSV Count	Area	Frequency	Throughput	Compute Density	Energy Efficiency	Accuracy
SRAM 2D	N/A	N/A	16 nm	SRAM Digital	SRAM CIM	0	0	0.114 mm^2	200 MHz	1.52 TOPS	13.3 TOPS/mm^2	50.1 TOPS/W	95.8%
Hybrid 2D	40 nm	40 nm	40 nm	SRAM Digital	RRAM CIM	1024	0	0.544 mm^2	200 MHz	1.52 TOPS	2.8 TOPS/mm^2	60.6 TOPS/W	99.3%
3-Tier H3D	40 nm	16 nm	16 nm	SRAM Digital	RRAM CIM	1024	5120	0.091 mm^2	185 MHz	1.41 TOPS	15.5 TOPS/mm^2	60.6 TOPS/W	99.3%

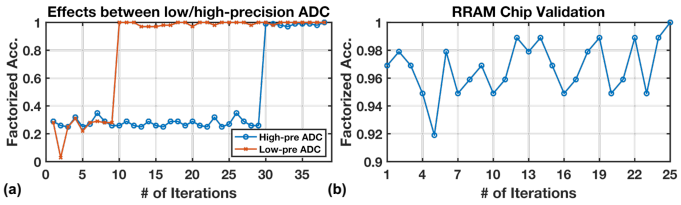


Fig. 6: Robustness Evaluation and Chip Validation. (a) Factorization accuracy with low-precision (H3DFACT) and high-precision ADC. (b) Factorization accuracy with 40 nm RRAM chip validation.

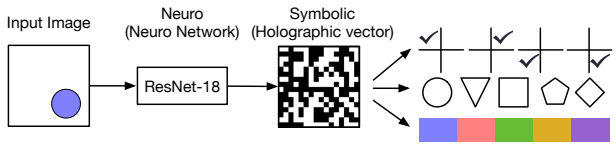


Fig. 7: Holographic Neuro-Symbolic Evaluation. The visual perception task involves neural networks for feature mapping and holographic vectors for attribute reasoning.

perceptual factorization with similar accuracy. As in Fig. 6a, after applying low-precision 4-bit ADC to similarity calculation, the factorization converges to 99% accuracy at 10th iteration, while it takes 30 iterations under 8-bit ADC. This is because lowering precision introduces quantization stochasticity, which prevents the factorizer stuck in a limit cycle and helps converge to the correct factorization in a shorter time (Fig. 2c).

RRAM Testchip Validation. We validate the effectiveness of our proposed H3DFACT on the fabricated 40 nm RRAM testchips [11], [13]. We extract inherent noise parameters from RRAM testchips by measuring the readout signal and incorporate their statistics into the developed holographic perceptual factorization framework. We also adjust the threshold value accordingly as the designed readout peripheral is able to change the readout voltage (V_{TGT} in Fig. 2). As in Fig. 6b, RRAM testchip validated H3DFACT achieves $> 96\%$ factorization accuracy at one-shot and reaches 99% accuracy after 25 iterations.

E. Holographic Perception Task Evaluation

Holographic Perception Accuracy. Fig. 7 demonstrates the role of H3DFACT in visual perception task to disentangle the attributes of raw images. The system consists of two components: a neural network to map input images to holographic perceptual vectors, and H3DFACT to disentangle the approximate product vector using a known set of image attributes (e.g., type, size, color, and position). Evaluated on the relational and analogical visual reasoning (RAVEN) dataset [25], H3DFACT achieves 99.4% accuracy of attributes estimation.

Extensible to Other Applications. H3DFACT is effective beyond visual perception, as factorization plays a fundamental role in perception and cognition (e.g., analogical reasoning, tree search, and integer factorization). We envision H3DFACT paves the way for solving complex combinatorial search problems in next-generation cognitive and neuro-symbolic AI systems.

VI. CONCLUSION

H3DFACT is the first H3D integrated CIM design unlocking efficient and scalable high-dimensional holographic vector factorization. H3DFACT exploits the computation-in-superposition capability and intrinsic hardware stochasticity, and consistently improves factorization accuracy and operational capacity, with 5.5× compute density, 1.2× energy efficiency, and 5.9× less silicon footprint compared to iso-capacity 2D designs. We envision H3DFACT being useful in exploring other robust and efficient cognitive and neuro-symbolic AI systems.

REFERENCES

[1] Y. Burak *et al.*, "Bayesian model of dynamic image stabilization in the visual system," *Proceedings of the National Academy of Sciences*, vol. 107, no. 45, pp. 19 525–19 530, 2010.
[2] M. Hersche *et al.*, "A neuro-vector-symbolic architecture for solving raven's progressive matrices," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 363–375, 2023.
[3] D. Kleyko *et al.*, "A survey on hyperdimensional computing aka vector symbolic architectures," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–52, 2023.
[4] D. Kleyko *et al.*, "Vector symbolic architectures as a computing framework for emerging hardware," *Proceedings of the IEEE*, vol. 110, no. 10, pp. 1538–1571, 2022.
[5] E. P. Frady *et al.*, "Resonator networks, 1: An efficient solution for factoring high-dimensional, distributed representations of data structures," *Neural Computation*, vol. 32, no. 12, pp. 2311–2331, 2020.
[6] A. Renner *et al.*, "Neuromorphic visual odometry with resonator networks," *arXiv preprint arXiv:2209.02000*, 2022.
[7] J. Langenegger *et al.*, "In-memory factorization of holographic perceptual representations," *Nature Nanotechnology*, vol. 18, no. 5, pp. 479–485, 2023.
[8] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive Computation*, vol. 1, pp. 139–159, 2009.
[9] G. Murali *et al.*, "On continuing dnn accelerator architecture scaling using tightly coupled compute-on-memory 3-d ics," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2023.
[10] S. Dutta *et al.*, "Monolithic 3d integration of high endurance multi-bit ferroelectric fet for accelerating compute-in-memory," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 36–4.
[11] S. D. Spetalnick *et al.*, "A 2.38 mcells/mm^2 9.81-350 tops/w rram compute-in-memory macro in 40nm cmos with hybrid offset/I_{OFF} cancellation and I_{CELL} R_{BLSL} drop mitigation," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2023, pp. 1–2.
[12] G. Karunaratne *et al.*, "Robust high-dimensional memory-augmented neural networks," *Nature communications*, vol. 12, no. 1, p. 2468, 2021.
[13] S. D. Spetalnick *et al.*, "A 40nm 64kb 26.56 tops/w 2.37 mb/mm 2 rram binary/compute-in-memory macro with 4.23 x improvement in density and $>$ 75% use of sensing dynamic range," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 1–3.
[14] F. Zhou *et al.*, "Near-sensor and in-sensor computing," *Nature Electronics*, vol. 3, no. 11, pp. 664–671, 2020.
[15] S. Yu *et al.*, "On the switching parameter variation of metal oxide rram—part ii: Model corroboration and device design strategy," *IEEE Transactions on Electron Devices*, vol. 59, no. 4, pp. 1183–1188, 2012.
[16] D. Bankman *et al.*, "An always-on $3.8\mu j/86\%$ cifar-10 mixed-signal binary cnn processor with all memory on chip in 28-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, 2018.
[17] W. Li *et al.*, "H3datten: Heterogeneous 3-d integrated hybrid analog and digital compute-in-memory accelerator for vision transformer self-attention," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2023.
[18] R. Swaminathan, "Advanced packaging: Enabling moore's law's next frontier through heterogeneous integration," in *IEEE Hot Chip Conference*, 2021, pp. 22–24.
[19] Y. Li *et al.*, "An adc-less rram-based computing-in-memory macro with binary cnn for efficient edge ai," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023.
[20] X. Peng *et al.*, "Dnn+ neurosim v2. 0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2020.
[21] X. Peng *et al.*, "Benchmarking monolithic 3d integration for compute-in-memory accelerators: overcoming adc bottlenecks and maintaining scalability to 7nm or beyond," in *2020 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2020, pp. 30–4.
[22] R. Mathur *et al.*, "Thermal-aware design space exploration of 3-d systolic ml accelerators," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 7, no. 1, pp. 70–78, 2021.
[23] UVA HotSpot, "Hotspot 6.0," https://lava.cs.virginia.edu/HotSpot/, 2019.
[24] Z. Fang *et al.*, "Temperature instability of resistive switching on hfox-based rram devices," *IEEE Electron Device Letters*, vol. 31, no. 5, pp. 476–478, 2010.
[25] C. Zhang *et al.*, "Raven: A dataset for relational and analogical visual reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 5317–5327.