H3DFact: Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations

Zishen Wan*, Che-Kai Liu*, Mohamed Ibrahim, Hanchen Yang, Samuel Spetalnick,
Tushar Krishna, Arijit Raychowdhury
School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA
{zishenwan, che-kai, mibrahim81, hanchen, sspetalnick3}@gatech.edu, {tushar, arijit.raychowdhury}@ece.gatech.edu

***Abstract*—Disentangling attributes of various sensory signals is central to human-like perception and reasoning and a critical task for higher-order cognitive and neuro-symbolic AI systems. An elegant approach to represent this intricate factorization is via high-dimensional holographic vectors drawing on 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, a 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 perceptual problems, H3DFact demonstrates superior capability in factorization accuracy and operational capacity by up to five orders of magnitude, with 5.5× compute density, 1.2× energy efficiency improvements, and 5.9× less silicon footprint compared to iso-capacity 2D designs.

I. Introduction

The brain's remarkable ability to reason and comprehend the world relies heavily on its capacity to disentangle sensory attributes. This intricate process involves the factorization of various sensory inputs (e.g., vision, hearing, touch) into distinct perceptual features. This factorization not only aids in perception but also serves as the foundation for higher-order cognition like problem-solving and abstract thinking, thus serving as a crucial component for neuro-symbolic AI [1]–[3].

An elegant approach to represent this intricate factorization is via high-dimensional holographic vectors in the context of brain-inspired vector-symbolic architecture [4]. Each sensory attribute is encoded and processed using a unique holographic vector, thereby creating distinct and separable representations. These representations can be manipulated using a set of algebraic operations. For instance, 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 when dealing with complex attribute structures [5].

The compositional nature of holographic vector representations gave rise to an efficient factorization algorithm, *resonator network*, that equips with superior ability to bridge cognitive gaps in neuro-symbolic AI, by accepting perceptual representations from a neural network and factorizing them for symbolic reasoning [6], [7]. 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 to circumvent the potential limit optimization cycle pitfalls. These features make factorization amenable to computing platforms that enable compute-in-memory (CIM) and are inherently stochastic, such as memristive devices [8].

Recently, an in-memory factorizer using the resonator network was proposed [9], 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 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 with the integration of multiple heterogeneous arrays in a 3D-stacked configuration [10].

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 improvements in silicon area and energy efficiency. Furthermore, the non-deterministic nature of the RRAM memory elements enhances factorization convergence and operational capacity. Compared to iso-capacity 2D designs, H3DFact demonstrates superior efficiency in terms of performance, power, 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 by virtue of inherent stochasticity, with 5.5× compute density, 1.2× energy efficiency, and 5.9× less silicon footprint compared to iso-capacity 2D designs.

[*] These authors contributed equally to this work.

979-8-3503-4859-0/DATE24/© 2024 EDAA

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 [11]: (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 illustrate 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 demonstrated 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):

$$
\begin{aligned}
\hat{x}(t+1) &= g\big(XX^\top(s \odot \hat{c}(t) \odot \hat{v}(t) \odot \hat{h}(t))\big); \quad X = [x_{cir}\ x_{tri}\ \ldots] \\
\hat{c}(t+1) &= g\big(CC^\top(s \odot \hat{x}(t) \odot \hat{v}(t) \odot \hat{h}(t))\big); \quad C = [c_{blue}\ c_{red}\ \ldots] \\
\hat{v}(t+1) &= g\big(VV^\top(s \odot \hat{c}(t) \odot \hat{x}(t) \odot \hat{h}(t))\big); \quad V = [v_{top}\ v_{bottom}] \\
\hat{h}(t+1) &= g\big(HH^\top(s \odot \hat{c}(t) \odot \hat{v}(t) \odot \hat{x}(t))\big); \quad H = [h_{left}\ h_{right}]
\end{aligned}
$$

where 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 around 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 [9]. One effective solution is to introduce stochasticity 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.

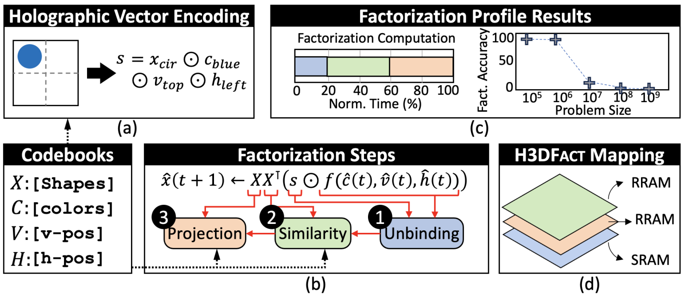


Fig. 1: Computational Primitives of the Holographic Vector Encoding and Factorization. (a) Vector encoding of a visual object. (b) Algorithmic flow of the factorization problem. (c) Characterization results of the factorization operations. (d) An overview schematic of the proposed H3D integrated factorizer with hybrid-memory design.

C. Heterogeneous 3D CIM Acceleration

Prior 3D integrated hardware designs have mainly focused on accelerating CNNs [12], Transformers [10], or monolithic 3D integration [13]. In contrast, H3DFact tackles a different MVM workload that is heavily used in high-dimensional cognitive systems, and maps different components of this factorization 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 [14]. It is worth noting that the existing array designs for VSAs [15] often fall short of fully supporting the bipolar space, as they frequently map a bipolar element $\{-1, +1\}$ to a single-bit quantity, which is not suitable for the factorization algorithm that 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

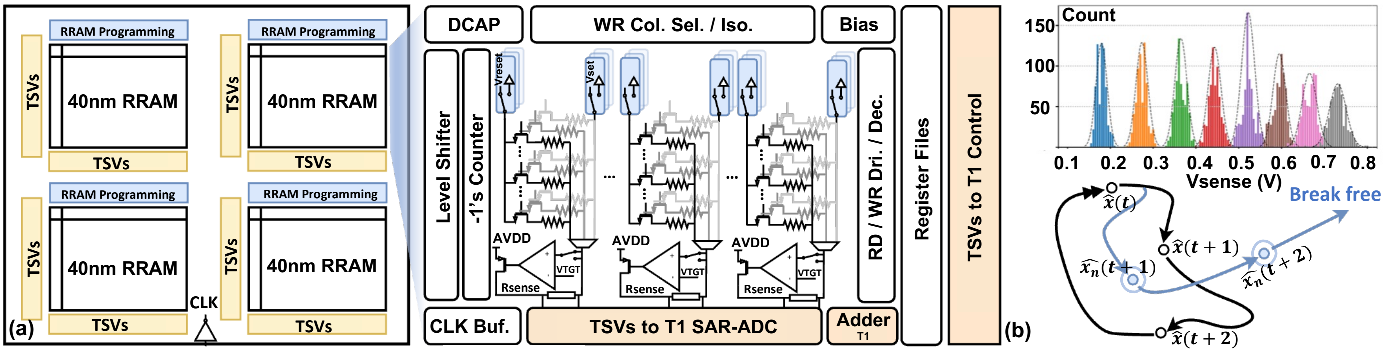


Fig. 2: H3DFact Array-Level Components. (a) Legacy node RRAM tier-level view and building blocks for a single RRAM array. (b) The inherent stochasticity of H3DFact helps break limit cycles and benefit factorization convergence.

(Fig. 2b). VTGT represents the target sensing voltage in the sensing path. Additionally, a current-sensing resistor (Rsense) is incorporated to enhance the process-voltage-temperature (PVT) 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 an RRAM chip fabricated with 40 nm technology [16].

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 (orange-colored blocks in Fig. 2a). One of the advantages of heterogeneous integration is the integration of systems with different technology nodes [17]. A potential area mismatch between RRAM and its peripherals results in MUX-sharing of the RRAM sensing [16]. To fully unleash the system performance, digital components in H3DFact are designed in 16 nm 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 RRAM is notorious for its humongous overhead [18]. The hybrid-computing scheme utilizes XNOR logic gates for bit-wise unbinding operation [19]. 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 through-silicon vias (TSVs) data transfer in the H3D design, which will be further explained in Sec. IV. To reduce the TSV overheads, 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 [10]. 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. Inspired by [9], one of the key insights 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.

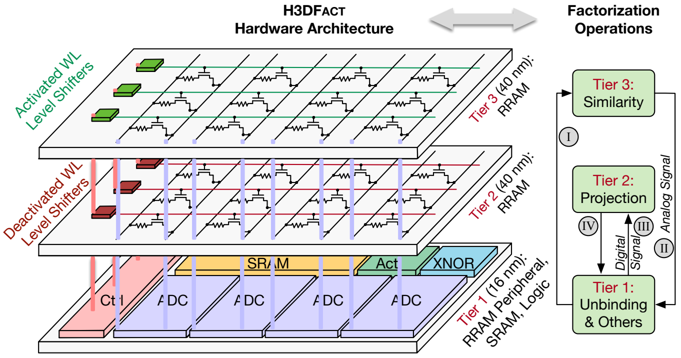


Fig. 3: H3DFact Architecture and Control Scheme. The factorization computation kernels are partitioned among three vertical tiers. The control scheme for activating only one tier of RRAM CIM arrays when all RRAM tiers share the same vertical interconnects. Turning off the power to WL level shifters (red) 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

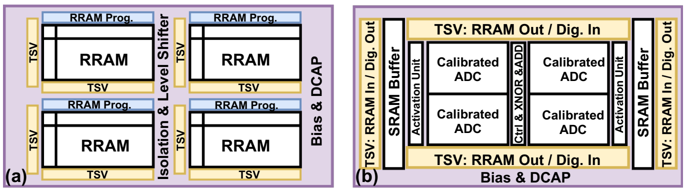


Fig. 4: H3DFact Floor Plan. (a) RRAM tier-2/3. (b) Digital tier-1.

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 the analog current can flow through the TSV one-shot. On the other hand, the 4-bit digital similarity results obtained from tier-1 are passed to projection tier-2 to avoid multi-bit digital value transmission degrading system performance Thus, H3DFact is designed with similarity at the top, projection at the middle, and digital circuits with advanced nodes at the bottom.

Tier-2 & Tier-3 RRAM CIM. Inspired by [10], a single set of RRAM peripherals is utilized collectively by both RRAM tiers (tier-2 and tier-3) in H3DFact, in a way that the interconnects from tier-1 link to every tier. Consequently, this architecture permits only one RRAM tier to be operational at any given time. This necessitates the inclusion of word line (WL) level shifters in each RRAM tier to manage their activation. Fig. 3 depicts the control scheme of two tiers of RRAM. Due to the shared peripherals, the RRAM WLs, bit lines (BLs), and source lines (SLs) across the tiers are effectively connected vertically. To ensure that only a single RRAM tier is activated at a time, the design is equipped with a full shutdown capability where the non-active RRAM cells do not contribute to 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 digital-SRAM in tier-1 to serve as buffers to support large batch factorization computation.

Design Methodology Generalization. The H3DFact architecture is adept at handling the diverse parameters characteristic of resonator networks. Since resonator networks are parametrized with high dimensional vector dimensions D and F (Sec. II), the H3DFact is configured with hardware dimensions to determine the number of rows of an RRAM array (d) and the number of RRAM subarrays of each tier (f). In this paper, we set $d = 256$ and $f = 4$ as an example of H3DFact design. This configuration not only accommodates the specified vector size, but also facilitates the parallel processing of multiple inputs by utilizing different subarrays.

B. *Tier-to-Tier Interconnects in* H3DFact

In Tab. I, we outline the parameters for the tier-to-tier interconnects in the H3DFact design. These assumptions are in line with recent H3D designs such as H3DAtten [10] and commercial designs such as AMD's 3D V-Cache [20]. For an RRAM array with X rows and Y columns, the total number of TSVs for connection to RRAM peripherals comprises X for WLs, Y for BLs, and $Y/2$ for SLs. Typically, larger arrays reduce TSV overhead but are less efficiently utilized than smaller ones. H3DFact opts to store the similarity matrix in the same array at each iteration to optimize TSV utilization.

Given the significant area costs associated with TSVs, our design strategy for H3DFact includes analog CIM and SAR-ADCs to minimize TSV area requirements. In analog CIM, the partial sums of MVM operations are conveyed as analog currents, requiring only a single set of interconnects to connect to an ADC in tier-1. H3DFact provides the flexibility to design RRAM peripheral circuitries in more advanced nodes [10], [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).

C. *Floor Planning and Bonding of* H3DFact

To verify that the tiers in the 3D stack of H3DFact are area-balanced and provide data for thermal analysis (Sec. V-C), we conduct a floor plan approximation for each tier. The sizes of the CIM arrays and their associated peripherals are estimated using the calibrated NeuroSim framework [22], which has been cross-validated with actual RRAM-based CIM macros [16]. Areas of other digital modules are extracted from the TSMC standard cell library.

Fig. 4a shows the floor plan of the H3DFact RRAM tier. Each RRAM subarray features a dimension of 256 × 256 size, with four subarrays designed for each tier. H3DFact can perform RRAM CIM operations 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 [12]. 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 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 [12], 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 perceptual systems. We demonstrate that H3DFact achieves improved factorization accuracy, operational capacity, and hardware efficiency, as well as illustrate the robustness of H3DFact with the RRAM silicon chip validation.

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
D=16	99.4	99.3	99.2	99.2	4	5	31	33
D=32	99.3	99.3	99.1	99.2	13	15	234	140
D=64	99.1	99.3	89.9	99.2	43	39	Fail	1347
D=128	96.9	99.3	0	99.2	Fail	108	Fail	17529
D=256	10.8	99.2	0	99.2	Fail	443	Fail	269931
D=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.

A. Accuracy and Operational Capacity

Accuracy Improvement. Tab. II compares the factorization accuracy of H3DFact with baseline resonator network [6] under different number of attributes F and code vectors D. While both baseline network and H3DFact achieve 99% accuracy under small M^D, H3DFact substantially enhances and maintains 99% accuracy under high dimensionality, illustrating its improved scalability for larger factorization problem sizes.

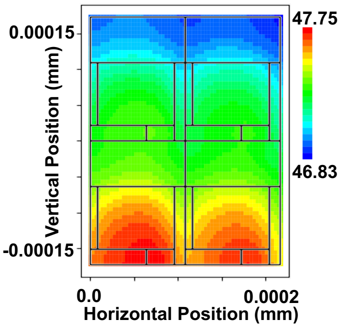
Operational Capacity Improvement. Tab. II also shows the number of iterations required to solve a given problem size with accuracy of at least 99%. Compared with baseline resonator network [6], H3DFact enables faster convergence and can solve problem sizes at five orders of magnitude larger at 99% accuracy, illustrating H3DFact capable of lowering computational complexity with improved operational capacity. This observation is in line with CIM-based factorizer design [9].

B. Hardware Efficiency

Monolithic 2D Baseline Design Setup. We evaluate the advantages of the proposed H3DFact by contrasting it with two distinct 2D architectures: a hybrid RRAM/SRAM design and an exclusively SRAM-based design (Tab. III). For the hybrid 2D design, all modules are integrated using a 40 nm process node to accommodate the RRAM technology in a unified 2D structure. Conversely, the fully SRAM design scales all modules to the more advanced 16 nm nodes. We maintain identical computing resources and parameters across all these designs to ensure an equitable comparison.

Silicon Footprint Reduction. Tab. III shows that fully SRAM design in 2D requires an area of 0.114 mm^2 with all components in 16 nm. The 2D RRAM/SRAM hybrid design occupies up to 0.544 mm^2 despite involving no TSV overheads due to limitations in current RRAM fabrication technology. In contrast, the advanced node scaling and vertical integration in H3DFact allow a more compact footprint of 0.091 mm^2. 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 marginally lower frequency due to the parasitic capacitance introduced by TSVs and hybrid bindings thus resulting in a slight throughput penalty. Nevertheless, as in Tab. III, H3DFact still demonstrates 1.2× higher compute density and 1.2× energy efficiency by scaling digital components and RRAM peripheral from 40 to 16 nm. Compared with the 2D fully deterministic digital SRAM baseline with all modules designed in 16 nm, H3DFact still achieves comparable energy efficiency with 5.5× higher compute density and 3.5% higher factorization accuracy due to the associated intrinsic stochasticity (Fig. 2c).



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/m^2°C)
Ambient temperature	25°C

Fig. 5: Thermal Analysis. Thermal map of H3DFact with its setup.

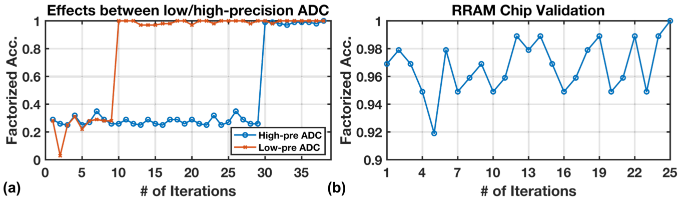


Fig. 6: Robustness Evaluation and Chip Validation. (a) Factorization accuracy with low-precision (H3DFact) and high-precision ADC. (b) Factorization accuracy with 40 nm RRAM chip validation.

Compare with Other Factorization Accelerators. Compared with recent PCM-based in-memory factorization [9], 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 Analysis. We utilized HotSpot [23] to conduct thermal analysis of H3DFact, assigning power densities to each component based on their respective floorplans (Fig. 4). Our chip-level thermal setup includes hybrid bonds and TSVs to connect tiers 1-3, C4 bumps to connect Tier 1 to the package, and thermal interface material (TIM) at the top for cooling. The parameters are summarized in Fig. 5 and consistent with [10]. 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 being more effective at the center and high power density lying in the southern of each macro, as expected, there exist slight temperature increases toward the die southern region. Importantly, the 3D stacking approach used in H3DFact does not compromise the reliability of RRAM, as RRAM retention is adversely affected at 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 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

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%

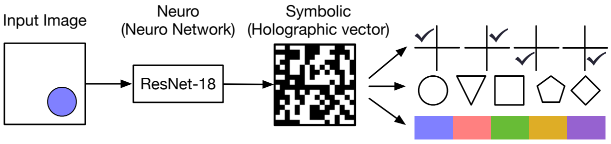


Fig. 7: Holographic Neuro-Symbolic Evaluation. The visual perception task involves neural networks for feature mapping and holographic vectors for attribute reasoning.

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 [14], [16]. 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.

ACKNOWLEDGEMENTS

The authors thank Wantong Li, Gauthaman Murali, and Lingjun Zhu for helpful discussions. This work was supported by CoCoSys, one of seven centers in JUMP2.0, a Semiconductor Research Corporation (SRC) sponsored by DARPA.

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] Z. Wan *et al.*, "Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai," *arXiv preprint arXiv:2401.01040*, 2024.

[4] D. Kleyko *et al.*, "A survey on hyperdimensional computing aka vector symbolic architectures," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–52, 2023.

[5] 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.

[6] 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.

[7] A. Renner *et al.*, "Neuromorphic visual odometry with resonator networks," *arXiv preprint arXiv:2209.02000*, 2022.

[8] S. Yu *et al.*, "Compute-in-memory: from device innovation to 3d system integration," in *ESSDERC 2021-IEEE 51st European Solid-State Device Research Conference (ESSDERC)*. IEEE, 2021, pp. 21–28.

[9] J. Langenegger *et al.*, "In-memory factorization of holographic perceptual representations," *Nature Nanotechnology*, vol. 18, no. 5, pp. 479–485, 2023.

[10] 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.

[11] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive Computation*, vol. 1, pp. 139–159, 2009.

[12] 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.

[13] 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.

[14] 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.

[15] G. Karunaratne *et al.*, "Robust high-dimensional memory-augmented neural networks," *Nature communications*, vol. 12, no. 1, p. 2468, 2021.

[16] 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.

[17] F. Zhou *et al.*, "Near-sensor and in-sensor computing," *Nature Electronics*, vol. 3, no. 11, pp. 664–671, 2020.

[18] 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.

[19] D. Bankman *et al.*, "An always-on 3.8μj/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.

[20] R. Swaminathan, "Advanced packaging: Enabling moore's law's next frontier through heterogeneous integration," in *IEEE Hot Chip Conference*, 2021, pp. 22–24.

[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] 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.

[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.