

CogSys: Efficient and Scalable Neurosymbolic Cognition System via Algorithm-Hardware Co-Design

Zishen Wan^{†*}, Hanchen Yang^{†*}, Ritik Raj^{†*}, Che-Kai Liu[†], Ananda Samajdar[‡],
Arijit Raychowdhury[†], Tushar Krishna[†]

[†]Georgia Institute of Technology, Atlanta, GA [‡]IBM Research, Yorktown Heights, NY
{zishenwan, hanchen, ritik.raj, che-kai}@gatech.edu, ananda.samajdar@ibm.com
{arijit.raychowdhury, tushar}@ece.gatech.edu

Abstract—Neurosymbolic AI is an emerging compositional paradigm that fuses neural learning with symbolic reasoning to enhance the transparency, interpretability, and trustworthiness of AI. It also exhibits higher data efficiency making it promising for edge deployments. Despite the algorithmic promises and demonstrations, unfortunately executing neurosymbolic workloads on current hardware (CPU/GPU/TPU) is challenging due to higher memory intensity, greater compute heterogeneity and access pattern irregularity, leading to severe hardware underutilization.

This work proposes CogSys, a characterization and co-design framework dedicated to neurosymbolic AI system acceleration, aiming to win both reasoning efficiency and scalability. **On the algorithm side**, CogSys proposes an *efficient factorization technique* to alleviate compute and memory overhead. **On the hardware side**, CogSys proposes a scalable neurosymbolic architecture with *reconfigurable neuro/symbolic processing elements (nsPE)* and *bubble streaming (BS) dataflow with spatial-temporal (ST) mapping* for highly parallel and efficient neurosymbolic computation. **On the system side**, CogSys features an *adaptive workload-aware scheduler (adSCH)* to orchestrate heterogeneous kernels and enhance resource utilization. Evaluated across cognitive workloads, CogSys enables reconfigurable support for neural and symbolic kernels and exhibits $>75\times$ speedup over TPU-like systolic array with only $<5\%$ area overhead, as benchmarked under the TSMC 28nm technology node. CogSys achieves $4\times$ – $96\times$ speedup compared to desktop and edge GPUs. For the first time, CogSys enables real-time *abduction reasoning* towards human fluid intelligence, requiring only 0.3 s per reasoning task with 4 mm^2 area and 1.48 W power consumption.

I. INTRODUCTION

The massive success of Large Language Models (LLMs) combined with concerns about interpretability and safety have led to an emerging paradigm of “compositional AI” systems - especially for safety-critical domains such as robotics and healthcare. The goal of such systems is to combine black-box neural networks with reasoning/rule-based AI methods [38], [43], [44], [49], [61], [87], [91], [98]. This approach mirrors human cognitive processes, which can be grouped as lower-level sensory processing (system 1 “thinking fast”) and higher-level cognitive functions like reasoning and deduction (system 2 “thinking slow”) [12], [17]. The former can be modeled with neural networks, and the latter with symbolic frameworks.

One promising example of compositional AI system is **neurosymbolic AI** that synergistically integrates neural network

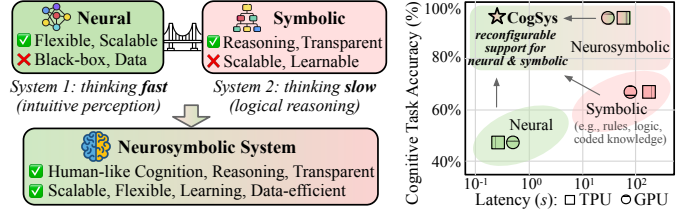


Fig. 1: Neurosymbolic AI is an emerging *compositional* system that integrates neural and symbolic modules, enabling superior cognitive intelligence compared to NNs. However, it suffers from inefficient TPU/GPU execution. **CogSys** is a reconfigurable neural/symbolic engine excelling in both reasoning efficiency and cognitive capability.

learning with *symbolic reasoning*. The neural networks are adept at identifying patterns and handling perceptual tasks, but lack transparency and logical inference capabilities. Symbolic modules (e.g., coded knowledge, rules), in contrast, excel in reasoning and interpretability but struggle with adaptability and learning from raw data. Neurosymbolic AI bridges this gap by composing strengths of both paradigms (Fig. 1).

Neurosymbolic AI has demonstrated superior capabilities in human-like reasoning and logical thinking across various domains, such as natural language processing, robotics, healthcare, etc [26], [29], [33], [35], [57], [59], [76], [87], [93], [96]. For example, IBM’s neuro-vector-symbolic system [33] achieves 98.8% accuracy on spatial-temporal reasoning tasks [95], greatly surpassing human (84.4%), ResNet (53.4%) and GPT-4 (89.0%); Google DeepMind’s Alpha-Geometry [83], another neurosymbolic system, solves geometry problems at a level of human Olympiad gold medalists, while GPT-4 completely fails. Recently, there also has been a plethora of workshops focusing on neurosymbolic AI [1]–[8].

Despite impressive cognitive capabilities of neurosymbolic AI - demonstrated by past work over distributed GPU clusters, recent study [86] identifies that enabling *real-time and efficient* neurosymbolic AI over edge devices, which is highly desirable for numerous reasoning and human-AI applications, is a challenging open problem. For example, a neuro-vector-symbolic system takes >3 mins even on TPU and desktop GPU for a single task [33]. This inefficiency threatens to hurt neurosymbolic AI deployment in the long run.

To understand this further, we systematically profile and analyze the runtime and memory behavior of various neu-

*Equal Contributions.

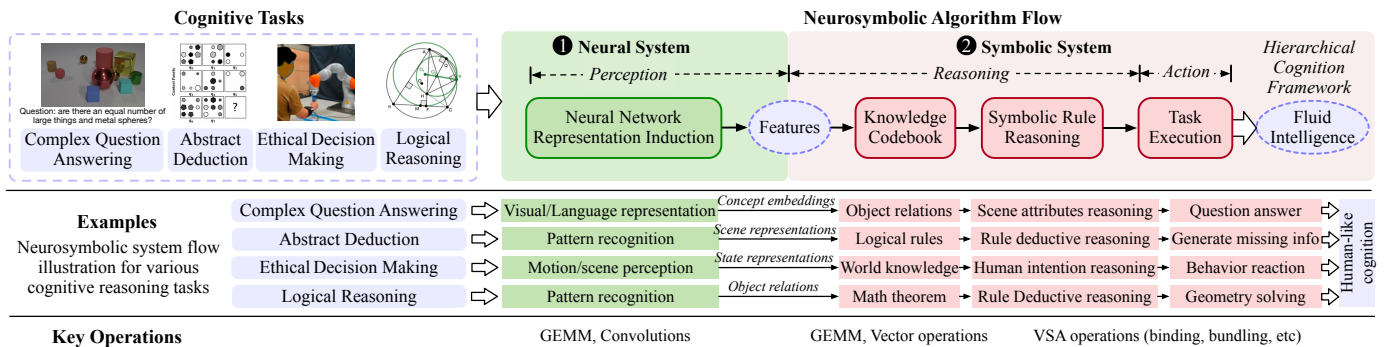


Fig. 2: Neurosymbolic algorithm flow. Neural systems handle perception by processing raw data and extracting features, which are then utilized by symbolic reasoning systems to apply logical rules and knowledge. This compositionality enables the execution of complex cognitive tasks such as abstract deduction, ethical decision-making, and fluid intelligence.

rosymbolic workloads on multiple devices and identify the following system-level challenges. **① Large Memory Footprint.** Neurosymbolic AI systems typically rely on vector-symbolic architecture (VSA) that utilizes vector operations to represent symbolic knowledge. The system generates an intermediate codebook that captures vast object combinations for higher reasoning capability (typically in the order of tens to hundreds MB) making it impractical to be cached on-chip in edge accelerators. **② Compute Heterogeneity.** Rather than general matrix multiplications (GEMMs) and convolution operations that current neural hardware largely focuses on accelerating, neurosymbolic workloads typically consist of numerous holographic vector operations (e.g., circular convolution) that run inefficiently on GPU and neural engines like TPUs due to low data reuse, low compute array utilization and low parallelism. **③ Sequential Processing.** Typically, the symbolic-reasoning computation depends on the output of the neuro-perceptual modules, increasing the critical path during cognitive inference and underutilizing parts of the accelerator.

To address the aforementioned challenges, we develop an algorithm-hardware co-design framework, dubbed CogSys, which to the best of our knowledge is the *first* to achieve real-time efficiency and scalability of cognitive neurosymbolic systems, making it more deployable and facilitate neurosymbolic AI development. **On the algorithm level,** CogSys proposes an *efficient factorization scheme* to reduce memory footprint. This technique completely replaces the large-size symbolic knowledge codebook, by quickly factorizing vectors in an interactive manner when decomposing symbolic representations. **On the hardware level,** CogSys proposes a scalable architecture with *reconfigurable neuro/symbolic processing elements (nsPE)* and *bubble streaming (BS) dataflow* with *spatial-temporal (ST) mapping* for highly parallel and energy-efficient neurosymbolic computation. The design is flexible to support heterogeneous neural and circular convolution symbolic kernels across vector dimensions and reduce runtime. **On the system level,** CogSys also features an *adaptive workload-aware scheduling (adSCH) scheme* with *multi-level parallelism* to orchestrate neural and symbolic kernels with improved hardware resource utilization and enables design scalability for evolving neurosymbolic AI. *Notably, with only <5% area overhead over TPU-like systolic*

arrays, CogSys enables reconfigurable support for neural and symbolic kernels and demonstrates >75× system speedup.

This paper, therefore, makes the following contributions:

- We perform comprehensive runtime and memory analysis of various neurosymbolic workloads across devices, and identify the primary cause of the inefficiency and optimization opportunities, which can also shed light on future neurosymbolic systems acceleration and innovations (Sec. III).
- We propose an algorithm-hardware co-design framework, dubbed CogSys, which is the first to enable real-time, efficient, and scalable VSA-based neurosymbolic systems, making it more deployable and facilitate neurosymbolic AI development.
- CogSys innovates across the algorithm-level efficient symbolic factorization strategy (Sec. IV), hardware-level reconfigurable neuro/symbolic architecture and dataflow (Sec. V), and system-level scheduler (Sec. VI) to reduce the memory footprint while improving hardware utilization and overall neurosymbolic processing efficiency.
- Evaluated across cognitive tasks, CogSys enables reconfigurable support for neural and symbolic operations, achieving 75.9× speedup with only a 4.8% area overhead compared to TPU-like systolic arrays, and demonstrates 4×-95× speedup compared to GPUs. CogSys enables efficient neurosymbolic AI with 4mm² area and 1.48W power consumption (Sec. VII).

II. NEUROSYMBOLIC AI BACKGROUND AND WORKLOAD

This section presents the preliminaries of neurosymbolic AI with its algorithm flow and key operations, then describes four representative neurosymbolic workloads for our analysis.

A. Challenges with Neural Networks

Neural methods are highly effective in extracting complex features from vision and language tasks, and excel in flexibility, scalability, and handling inconsistency [68], [88]–[90], [100]. However, neural methods often suffer from limitations such as hallucinations and lack of interpretability, and typically operate as black-box where their decision-making processes are not easily understandable by humans. This undermines the model output trustworthiness in cognitive and safety-critical applications [31], [39], [84].

B. Neurosymbolic AI Algorithm Flow

Neurosymbolic AI synergistically integrates the learning capability of neural networks with the reasoning capability of symbolic AI, offering advantages in data-efficient learning with transparent and logical decision-making compared to DNNs. Neurosymbolic AI leads to superior performance in a wide range of applications, such as complex question answering [57], [59], abstract deduction [33], [96], decision making [64], [78], logical reasoning [70], [83] tasks, serving as a promising paradigm to achieve human-like fluid intelligence.

Fig. 2 extracts a unified neurosymbolic pipeline and illustrates how they interact to perform complex cognitive tasks:

① **Neural system.** The process begins with the neural module handling perception tasks by interpreting sensory data and generating meaningful scene and object representations which are essential for further reasoning processes. The neural module itself may suffer from superposition catastrophe, preventing it from extracting object constituent attributes [33].

② **Symbolic system.** The extracted features are fed into the symbolic system for reasoning tasks. This step enhances explainability and reduces dependence on extensive training data by incorporating established models of the physical world (e.g., underlying rules, coded knowledge). Throughout this process, a knowledge codebook is maintained, which integrates learned knowledge from the neural network with symbolic rules, ensuring that the system can both learn from new data and apply logical reasoning based on existing knowledge. The outcomes of symbolic reasoning process are then used to make decisions, generates responses, or controls actions.

This neurosymbolic flow is one way to model human hierarchical reasoning procedures. Resembling the sense-reason-act cognitive cycle can be computationally modeled through a multi-layer framework [37], [65], where *perception* layer fuses sensory inputs and maps them to high-level observations, *reasoning* layer conducts deliberate and conscious thinking by applying symbolic rules and knowledge, *action* layer facilitates trustworthy and reliable execution. This compositional approach allows agents to tackle complex challenges that require both data-driven learning and logical reasoning.

C. VSA-Based Symbolic Operations

Vector-symbolic architecture (VSA). Within the compositional neurosymbolic AI flow, exploiting VSA with neural dynamics has become the powerful approach [24], [32], [33], [46], [60], [96]. Specifically, VSA provides a means to represent symbolic information in a low or high-dimensional vector space. By encoding symbolic structures as vectors with dimensionality-preserving algebraic operations, VSAs enable the combination of symbolic reasoning with neural networks, thereby facilitating cognitive tasks such as learning, memory, and reasoning in a unified system. Fig. 3 illustrates a simple example of the binding ambiguity of neural networks and the functionality of VSA structures.

Circular convolution. A key VSA operation is the block-wise *circular convolution*, serving as a universal operation for vectors representing different symbols. Circular convolution

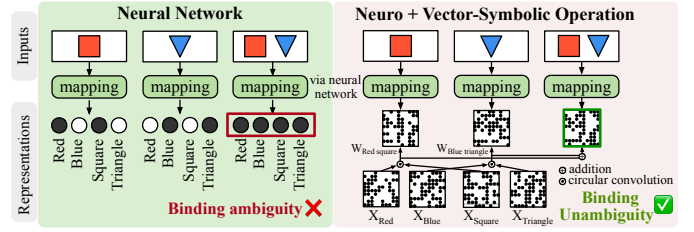


Fig. 3: Illustration of VSA functionality. Neural network suffers from binding ambiguity issues, whereas VSA constructs vector representations with circular convolution operations for reasoning process.

combines two vectors in a way that preserves the information from both, making it suitable for representing composite symbols. Mathematically, the circular convolution of two vectors \mathbf{A} and \mathbf{B} (each of dimension N) generates vector \mathbf{C} as $C[n] = \sum_{k=0}^{N-1} A[k] \cdot B[(n-k) \bmod N]$ where each element of \mathbf{C} is obtained by multiplying the elements of \mathbf{A} with the circularly shifted elements of \mathbf{B} , and then summing up. This process is repeated for each element n (0 to $N-1$). Circular convolution has commutativity and associativity properties, making it particularly effective in hierarchical reasoning tasks where manipulating structured information is critical.

Symbolic knowledge codebook. Symbolic knowledge is typically represented as a set of codebooks for the attributes of interest (Fig. 3). To describe an object with various attributes, a product vector can be computed by binding knowledge codebooks via circular convolution. Due to the properties of multiplicative binding, the co-activated VSA representations result in minimal interference, allowing each object to be accurately recovered. The query vector generated from neural networks will be compared with all codebook vectors to derive attributes for further reasoning. The codebook is typically in the order of tens to hundreds of MB, making it impractical to be cached on-chip in edge accelerators for complex tasks.

D. Representative Neurosymbolic AI Models

Following the flow in Fig.2, we analyze four VSA-based neurosymbolic workloads in detail: NVSA [33] for spatial-temporal reasoning, MIMONet for multi-input simultaneous processing [60], LVRF for probabilistic abduction [32], and PrAE for abstract reasoning [96]. These workloads achieve state-of-the-art performance and unlock advanced reasoning capabilities. Our goal is to understand their system and architectural challenges to enable scalable neurosymbolic deployment, where latency and efficiency are critical factors.

Tab. I lists the details of selected representative workloads:

1) *Neuro-Vector-Symbolic Architecture (NVSA)*: NVSA is a neurosymbolic system advancing spatial-temporal abduction reasoning [33]. Its *neural* module handles visual perception, while the *symbolic* module uses VSA-based operations for probabilistic inference, symbolic rule reasoning, and execution. NVSA bypasses the superposition catastrophe [67] and surpasses neural-only methods, achieving human-level performance on key fluid intelligence reasoning tests [95].

2) *Multiple-Input-Multiple-Output Networks (MIMONet)*: MIMONet is a neurosymbolic model designed to handle

TABLE I: Neurosymbolic models. Selected neurosymbolic AI workloads for analysis, representing a diverse of application scenarios.

Representative Neuro-Symbolic AI Workloads	Neuro-Vector-Symbolic Architecture [33]	Multiple-Input-Multiple-Output Neural Networks [60]	Probabilistic Abduction via Learning Rules in Vector-symbolic Architecture [32]	Probabilistic Abduction and Execution Learner [96]
Abbreviation	NVSA	MIMONet	LVRF	PrAE
Learning Approach	Supervised/Unsupervised	Supervised	Supervised	Supervised/Unsupervised
Compute Pattern	Neuro: CNN Symbolic: VSA binding/unbinding (Circular Conv)	CNN/Transformer VSA binding (Circular Conv)	Supervised VSA binding/unbinding (Circular Conv)	Supervised/Unsupervised Probabilistic abduction
Application Scenario	Use Case : Spatial-temporal reasoning, Fluid intelligence, Abstract reasoning Advantage vs. Neural Model : Higher joint representation efficiency, Better reasoning capability, Transparency	Multi-input simultaneously processing with single CNN/Transformer Higher throughput, Lower latency, Compositional compute, Transparency	Probabilistic reasoning, Analogy reasoning, Out-of-distribution (OOD) data processing Stronger OOD handling capability, One-pass learning, Higher flexibility, Transparency	Spatial-temporal reasoning, Fluid intelligence, Abstract reasoning Higher generalization, Transparency, Interpretability, Robustness

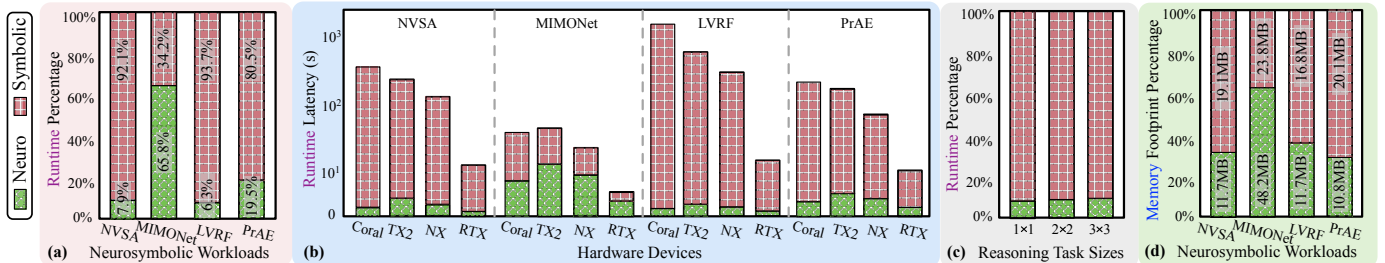


Fig. 4: End-to-end neurosymbolic runtime, memory, and roofline characterization. (a) Benchmark neurosymbolic models on CPU+GPU system, showing symbolic may serve as system bottleneck. (b) Benchmark neurosymbolic models on Coral TPU, TX2, NX, and 2080Ti GPU, showing that real-time performance cannot be satisfied. (c) Benchmark models on various task sizes, indicating the potential scalability problem. (d) Benchmark memory footprint of neurosymbolic models, showing large memory overhead of symbolic knowledge codebook.

multiple inputs and reduce computational cost per input [60]. Its *neural* modules use CNN/Transformer architectures, while its *symbolic* modules employ VSA binding/unbinding for encoding/decoding, enabling computation in superposition. MIMONet achieves 2-4 \times speedup with higher accuracy on LRA benchmarks compared to neural-only methods [82].

3) *Probabilistic Abduction via Learning Rules in Vector-symbolic Architectures (LVRF)*: LVRF is a neurosymbolic architecture for visual reasoning and handling out-of-distribution data [32]. Its *neural* modules handle visual perception, while *symbolic* modules use VSA for probabilistic abduction reasoning. LVRF outperforms neural-only methods in unseen reasoning tasks [36], offering greater flexibility and interpretability.

4) *Probabilistic Abduction and Execution (PrAE) Learner*: PrAE is a neurosymbolic learner for spatial-temporal cognitive reasoning [96]. Its *neural* modules handle visual perception and produce scene representations, while the *symbolic* modules conduct probabilistic reasoning and abduct rules. PrAE offers human-level generalizability, transparency, and interpretability, which classic neural networks struggle to achieve.

III. NEUROSYMBOLIC AI SYSTEM PROFILING

Building upon prior profiling study [87], this section characterizes the system behavior of various vector-symbolic-based neurosymbolic models (Sec. III-A-III-D), and provides workload insights for computer architects (Sec. III-E, III-F).

A. Profiling Setup

To understand the real-device efficiency of neurosymbolic AI workload, we profile four representative models as elaborated in Sec. II-D, in terms of runtime, memory, and compute operators, for solving cognitive reasoning problems on four devices, including Coral edge TPU (4 W), Jetson TX2 (15 W), Xavier NX (20 W), and RTX 2080Ti (250 W), respectively.

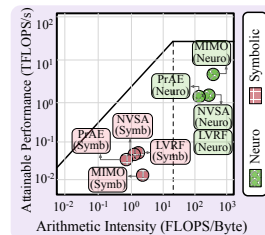


Fig. 5: Roofline analysis. End-to-end neurosymbolic roofline characterization on RTX 2080Ti GPU, indicating that typically neuro is compute-bound and symbolic is memory-bound.

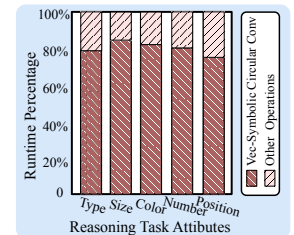


Fig. 6: Symbolic operation analysis. Symbolic operations are dominated by vector-symbolic circular convolution and vector-vector multiplication stemming from hypervector representations.

B. Compute Latency Analysis

End-to-end latency breakdown. Fig. 4a and Fig. 4b illustrate the end-to-end latency breakdown of four neurosymbolic workloads. We can observe that (1) *The real-time performance cannot be satisfied* on all four devices. Even if more computing resources are available to reduce NN runtime, the significant overhead of symbolic reasoning still prohibits real-time execution. (2) *Symbolic operations consistently dominate runtime.* For example, the symbolic modules count for 87% of total NVSA inference time while its floating-point operations (FLOPS) count for only 19% of total NVSA FLOPS, indicating that the symbolic operations may not be well executed by GPU/TPU. (3) *Symbolic reasoning computation lies on the critical path* due to its dependence on the neuro workloads.

End-to-end latency scalability. Fig. 4c indicates that the neuro vs. symbolic runtime proportion remains relatively stable across various tasks and sizes. For example, when Raven’s Progressive Matrices (RPM) [95] task size increases from 2 \times 2 to 3 \times 3, the NVSA symbolic modules runtime changes from 91.6% to 87.4%, while the total runtime increases by 5.02 \times on average across 14 test scenarios, *indicating the scalability*

TABLE II: Hardware inefficiency analysis. The compute, memory, and communication characteristics of representative neural and symbolic kernels on CPU/GPU platform.

	Neural Kernel		Symbolic Kernel	
	sgemm_nn	relu_nn	vectorized_elem	elementwise
Compute Throughput (%)	95.1	92.9	3.0	2.3
ALU Utilization (%)	90.1	48.3	5.9	4.5
L1 Cache Throughput (%)	79.7	82.6	28.4	10.8
L2 Cache Throughput (%)	19.2	17.5	29.8	22.8
L1 Cache Hit Rate (%)	1.6	51.6	29.5	33.3
L2 Cache Hit Rate (%)	86.8	65.5	48.6	34.3
DRAM BW Utilization (%)	14.9	24.2	90.9	78.4

bottleneck of neurosymbolic models.

C. Memory and System Analysis

Memory footprint. Fig. 4d characterizes the memory footprint of neurosymbolic workloads. We can observe that (1) *Neural weights and symbolic codebooks typically consume large storage footprint*, because neural perception enables the expression of more object combinations than vector space dimensions, requiring the codebook to be large enough to ensure vector quasi-orthogonality. (2) Symbolic modules consume large memory due to a large number of vector operations depending on intermediate results and exhaustive search.

System Roofline Analysis. Fig. 5 employs the roofline model of RTX 2080Ti GPU version to quantify the neurosymbolic workloads. We observe that *symbolic modules are memory-bounded while neuro modules are compute-bounded*. This is mainly due to symbolic operations requiring streaming vector elements, increasing the memory bandwidth pressure and resulting in hardware underutilization (Sec. III-D).

D. Symbolic Operation and Inefficiency Analysis

Symbolic operation analysis. Inspired by the dominated symbolic bottleneck, we analyze their detailed operations in Fig. 6. We observe that vector-symbolic *circular convolution and vector-vector multiplication dominate symbolic modules*, accounting for 80% of runtime. In contrast to the shared neural modules, these symbolic operations run sequentially and separately for each downstream cognition task and underlying rule. These operations typically stem from high-dimensional distributed vectors and are difficult to process efficiently on

GPU/TPU. Thus, the challenges of accelerating these vector-symbolic computations will become increasingly important as the cognitive task and feature complexities further grow.

Symbolic hardware inefficiency analysis. To further quantify the reason for hardware inefficiency, we analyze the compute and memory units behavior of representative neuro and symbolic kernels, as listed in Tab. II. *The system inefficiencies mainly come from ALU underutilization, low cache hit rate, and massive data movement of symbolic operations*. Symbolic data transfer accounts for half of total latency, where >80% is from host to GPU, while neural kernels exhibit high utilization.

E. Unique Characteristics of Neurosymbolic vs ML Workloads

To summarize, based on above characterization, neurosymbolic AI differs from ML workloads mainly in three aspects:

Compute kernels. Neurosymbolic workloads consist of heterogeneous neural and symbolic kernels. Symbolic operations execute inefficiently on CPU/GPU/TPU with low hardware utilization and cache hit rate, resulting in latency bottleneck.

Memory. Symbolic operations are memory-bounded due to large element streaming for vector-symbolic operations. Symbolic codebooks typically account for large memory footprints and require large intermediate caching during computation.

Dataflow and scalability. Neurosymbolic workloads exhibit more complex control than CNNs. Symbolic modules typically have irregular dataflow, data dependency, and sequential processing, bringing low parallelism scalability and inefficiency.

F. Identified Opportunities for Neurosymbolic Optimization

While neurosymbolic AI holds great promise, addressing its inefficiencies is critical for achieving real-time, scalable deployment and ensuring long-term development. To this end, we propose CogSys, an algorithm-hardware co-design framework designed to enhance both reasoning energy efficiency and accuracy in neurosymbolic AI (Fig. 7). **At the algorithm level**, hardware-friendly codebook optimization reduces memory footprint and latency (Sec. IV). **At the hardware level**, the architecture and dataflow must be efficient for VSA operations and reconfigurable for neuro/symbolic kernels (Sec. V). **At the system level**, the architecture must efficiently and adaptively schedule diverse neurosymbolic workloads (Sec. VI). CogSys

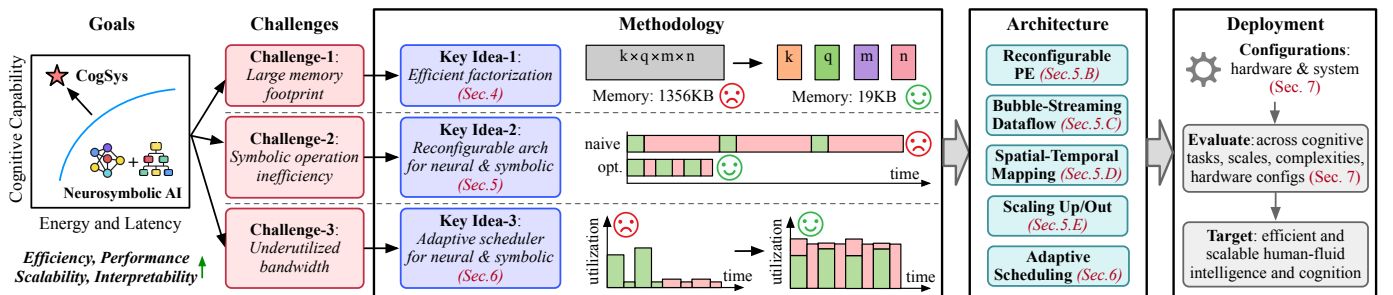


Fig. 7: CogSys system overview. CogSys is an algorithm-hardware co-design framework for neurosymbolic AI with the goal to achieve efficient and scalable human-fluid intelligence and cognition systems. CogSys addresses the challenges of redundant storage, symbolic and vector operation latency bottleneck, sequential processing and hardware underutilization, by proposing methodologies including efficient factorization, reconfigurable PE, efficient dataflow, mapping, scalable architecture, multi-level parallelism and scheduler. CogSys is deployed across cognitive applications and consistently demonstrates performance-efficiency-accuracy improvements for neurosymbolic systems.

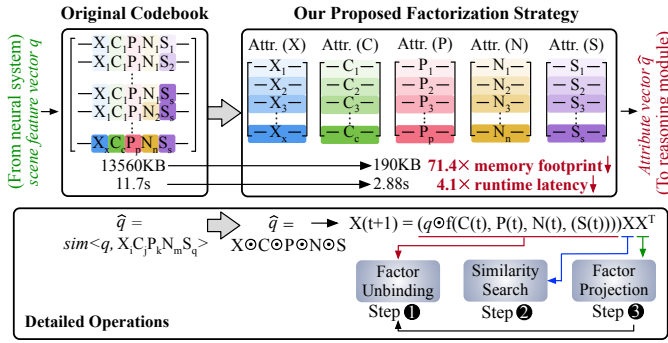


Fig. 8: Proposed symbolic codebook factorization strategy. The efficient factorization technique quickly factorizes a product vector in an interactive manner and significantly reduces the memory footprint demand of the symbolic codebook when decomposing query vectors.

consistently demonstrates improvements in performance, efficiency, and accuracy across reasoning applications (Sec. VIII).

IV. COGSYS: ALGORITHM OPTIMIZATION

This section presents our proposed CogSys algorithm optimizations for efficient and scalable neurosymbolic systems. We propose an *efficient vector-symbolic factorization strategy* to reduce the large memory footprint (Sec. IV-A), and explore the *stochasticity injection* and *low-precision operation* to accelerate neurosymbolic systems (Sec. IV-B).

A. Symbolic Factorization Strategy

Overall pipeline. To address the large memory footprint of symbolic codebooks (Sec. III-C), we propose an efficient factorization strategy. The key idea is to disentangle the large volume of object combination vectors in symbolic knowledge codebook into the small volume of basis attribute vectors, thus lowering computational time and space complexity (Fig. 8).

Specifically, given an entangled query vector q (e.g., scene representation) generated from neural module and the set of symbolic codebooks $\{X^i\}_{i=1}^F$ (each with M possible solutions and F codebooks in total), this creates a combinatorial search and storage involving M^F vectors. Instead of searching along all possible combinations, our factorization method iteratively searches in superposition to find the valid $\hat{x}^i \in X^i$ such that the estimated vector $\hat{q} = \hat{x}^1 \odot \hat{x}^2 \odot \dots \odot \hat{x}^f$ resembles with the highest similarity to the input query q . By exploiting the quasi-orthogonality of the vectors, our factorization module is able to rapidly search through the various combinations in superposition by iteratively unbinding all but one of the factors from the product vector, and then projecting it into the space of possible solutions of the considered factor that is used for the following reasoning procedure. In this way, we can replace the original symbolic codebook and greatly reduce its storage.

Detailed steps. Fig. 8 illustrates our symbolic knowledge codebook factorization strategy, consisting of three steps:

Step 1: Factor unbinding via element-wise multiplication (\odot). For a given factor, the unbinding is performed by taking the product vector q and unbinding the contribution of the other factors' latest estimate: $\tilde{x}^i(t) = q \odot \prod_{f=1}^F \hat{x}^f(t)$ ($f \neq i$).

TABLE III: Algorithm optimization impact. Factorization, stochasticity, and quantization impact accuracy, latency, and memory.

	Accuracy (higher is better)	Latency (lower is better)	Memory (lower is better)
Factorization	Increase	Reduce	Reduce
Stochasticity	Increase / Reduce	Reduce	No Impact
Low-Precision	Reduce	Reduce	Reduce

Step 2: Similarity search via matrix-vector multiplication. The similarity vector $\alpha^f(t)$ is calculated for each unbound estimate: $\alpha^f(t) = \tilde{x}^f(t) \cdot X^f$, $\forall f \in [1, F]$.

Step 3: Factor projection via matrix-vector multiplication. The estimates for the factors for the subsequent time step are given by the linear combination of all the codevectors with the similarity vectors acting as weights: $\hat{x}^f(t+1) = \text{sign}(\alpha^f(t) \cdot (X^f)^T)$, $\forall f \in [1, F]$. Repeat Steps 1-3 until convergence.

Applicable across neurosymbolic workloads. Our proposed efficient factorization module can apply to various levels of conceptual hierarchy, such as factoring time-varying pixel data of dynamic scenes [9], factoring sentence structure into roles and fillers [53], and cognitive analogical reasoning [33]. Essentially, given its ubiquitous applicability in perception and cognitive reasoning, we envision it being a core component in future large-scale neurosymbolic cognitive systems.

B. Stochasticity and Low-Precision Operation

Factorization optimization via stochasticity. To reduce the required number of iterations of factorization, we propose to apply additive Gaussian noise. We observe that injecting stochasticity in both Step 2 similarity and Step 3 projection operations helps the factorization process escape limit cycles, allowing it to explore a much larger solution space and achieve faster convergence (Tab. VIII).

Operator precision optimization. To further reduce the memory footprint, we apply quantization techniques to the workloads. Specifically, we apply 8-bit floating-point and integer arithmetic for both neuro and symbolic computations with fine-tuning to maintain reasoning accuracy (Tab. IX).

C. Algorithm Optimizations Discussion

Impact on accuracy, latency, and memory. The factorization, stochasticity, and quantization optimizations impact accuracy, latency, and memory requirements to varying extents. As shown in Tab. III, accuracy improves with factorization (due to precise attribute extraction, aiding downstream symbolic reasoning) and with stochasticity optimizations (due to improved convergence). However, quantization results in accuracy decreases due to data imprecision. Designers can balance speed and accuracy by tuning factorization convergence threshold.

V. COGSYS: HARDWARE ARCHITECTURE

This section presents CogSys architecture, the *first* hardware to enable efficient and scalable neurosymbolic processing. CogSys architecture features *reconfigurable neuro/symbolic processing elements (nsPE)* (Sec. V-B), *bubble streaming (BS) dataflow* (Sec. V-C), *adaptive spatial-temporal (ST) mapping* (Sec. V-D) with scalable array architecture (Sec. V-E) and customized SIMD units (Sec. V-F) for neurosymbolic processing.

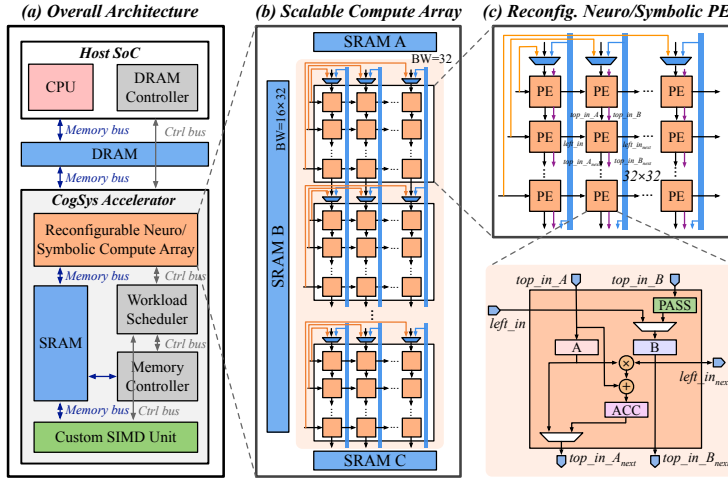


Fig. 9: Architecture overview. CogSys system includes DRAM, host System-on-Chip (SoC), and CogSys accelerator that consists of five major components: reconfigurable and scalable neuro/symbolic compute arrays, custom SIMD units, double-buffered SRAMs, workload scheduler, and memory controller.

TABLE IV: Comparison between CogSys with other accelerators. CogSys is the first to enable efficient and scalable vector-symbolic circular convolution (CircConv) and neurosymbolic models.

Accelerators	Reconfigurable	Scale-up/Scale-out	Memory Footprint for Single CircConv (Vector Dimension = d)	CWP* Support for CircConv	ScWP** Support for CircConv	Neuro-symbolic AI Support
Eyeriss [15]	No	Scale-up	No Support	No	No	No
Neuro Cube [45]	No	Scale-out	$O(d^2)$, GEMV	No	Yes	No
Brainwave [16]	No	Scale-out	$O(d^2)$, GEMV	No	Yes	No
SARA [72]	No	Both	$O(d^2)$, GEMV	No	Yes	No
TPU [42]	No	Scale-out	$O(d^2)$, GEMV	No	Yes	No
SIMBA [77]	No	Scale-out	$O(d^2)$, GEMV	No	Yes	No
MTIA [19]	No	Scale-out	$O(d^2)$, GEMV	No	Yes	No
CogSys (Ours)	Yes	Both	$O(d)$, BS Dataflow	Yes	Yes	Yes

*Column-Wise Parallelism (within a systolic cell); **Systolic cell-Wise Parallelism

A. Overview of Proposed CogSys Architecture

Neurosymbolic workloads feature much greater heterogeneity in compute kernels than DNNs, leading to an increasing divergence with the current hardware that focuses on GEMMs and convolutions. As illustrated in Tab. IV, CogSys is proposed, for the first time, to support neurosymbolic workloads and achieve efficient and scalable implementation of symbolic operations.

Aiming to design a complete neurosymbolic acceleration system, our design includes DRAM, a host SoC, and CogSys accelerator consisting of five major components: reconfigurable neuro/symbolic compute array, SIMD unit, double-buffered SRAM, adaptive scheduler and memory controller (Fig. 9). During the reasoning procedure, the host SoC streams task in, and then the reconfigurable arrays perform neuro (e.g., GEMMs/convolutions) and vector-symbolic operations (e.g., circular convolution), while the SIMD units accelerate element- and vector-wise operations with multi-level parallelism and adaptive workload scheduling. *It is worth noting that monolithic systolic array (TPU-like) is extremely inefficient for symbolic workloads (Sec. V-C), while CogSys provides reconfigurable support for neural and symbolic kernels and demonstrates $>75\times$ speedup with only $<5\%$ area overhead over systolic array architecture.*

Reconfig. Neuro/Symbolic PE Operation Mode

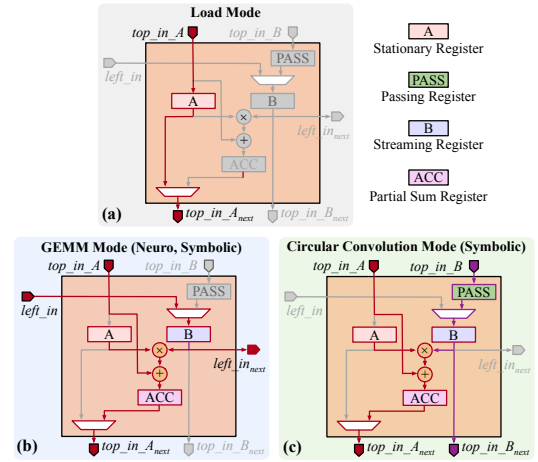


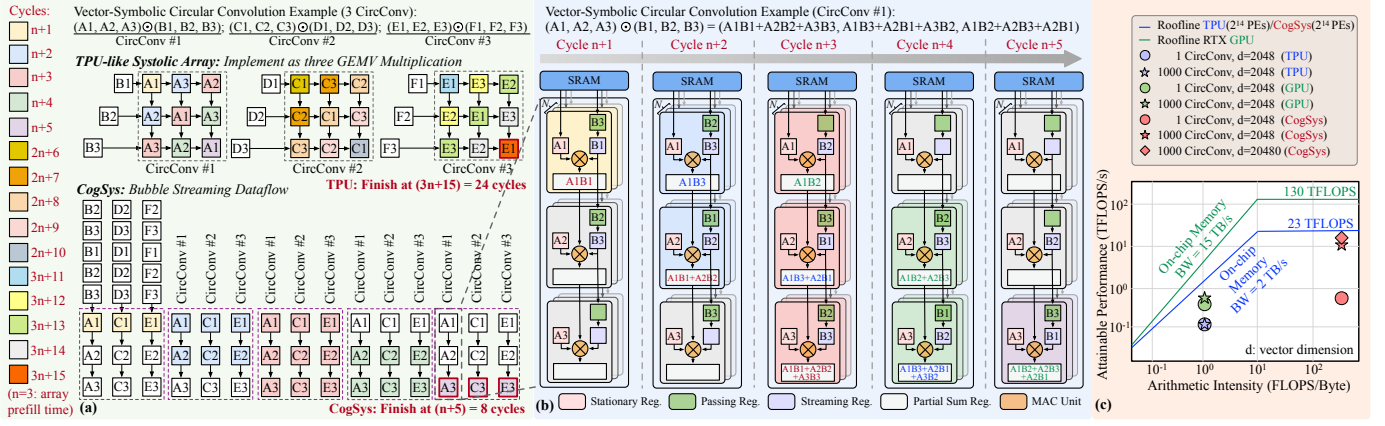
Fig. 10: Reconfigurable neuro/symbolic PE ($nsPE$). Each $nsPE$ includes four registers and supports three modes (load, neuro, symbolic) that provide reconfigurable support for neurosymbolic operations.

B. Reconfigurable Neuro/Symbolic Processing Element

Reconfigurable neuro/symbolic PE ($nsPE$). Instead of having separate PEs for neuro and symbolic operations that incur large area overhead, we propose $nsPE$ micro-architecture that provides reconfigurable support to both neuro and symbolic operations (Fig. 10). Each $nsPE$ consists of four registers (stationary, passing, streaming, and partial sum registers) and supports three operation modes (load, GEMM, and circular convolution). During load mode, the input vectors A (weights of GEMM) are passed into the stationary register using ‘top_in_A’ links. Reconfigurability is achieved by selecting input B either from ‘left_in’ link (GEMM mode) or the passing register (circular convolution mode). During GEMM mode, the $nsPE$ operates as TPU-like architecture for efficient GEMM and convolution. Inputs are streamed from left to right using ‘left_in’ links. During circular convolution mode, input vector B is streamed from top to bottom using ‘top_in_B’ links with a bubble via passing register (Sec. V-C), facilitating the temporal reuse of the streaming input for efficient vector-symbolic circular convolution operation. The reconfigurable $nsPE$ can also support efficient circular correlation by reversing stationary vector A . During both GEMM and circular convolution modes, partial products are reduced from top to bottom with ‘top_in_A’ links.

C. Efficient Bubble Streaming Dataflow

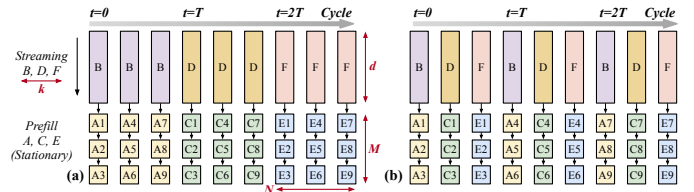
Inefficiency of TPU-like systolic array. TPU-like systolic array (SA) exhibits high memory footprint and low parallelism for symbolic circular convolution operations. Fig. 11a shows a scenario of three circular convolutions. TPU-like systolic cell implements them as general matrix-vector (GEMV) multiplication where matrices contain circularly shifted stationary vectors with the matrix memory footprint of $O(d^2)$. Additionally, TPU-like SA is incapable of parallelizing multiple GEMV on a systolic cell and need to process them sequentially.



Bubble streaming (BS) dataflow. To efficiently support symbolic operations in $nsPE$ array, we propose BS dataflow for circular convolution (Sec. II-C) and circular correlation (opposite direction circular shift) which is the vector-symbolic bottleneck. Fig. 11b presents an example of BS dataflow performing circular convolution of two vectors A and B ($d=3$) on a 3×1 $nsPE$ array. In BS dataflow, vector B is streamed from one $nsPE$ to another through bubbles (passing registers) while vector A is held in stationary registers. The BS dataflow enables a passing register to temporarily store the streaming input for a cycle before it moves to the streaming register. This value is transferred to the passing register of the next $nsPE$ in the following cycle. The MAC unit processes the data from stationary and streaming registers, adding it to the partial product. The procedure is repeated until final outputs.

Improved arithmetic intensity of BS dataflow. The BS dataflow achieves higher arithmetic intensity than GEMV in GPU/TPU-like systolic cells, as illustrated in roofline analysis (Fig. 11c). This efficiency mainly comes from reduced memory footprint and increased parallelism (Tab. IV). (1) Linear memory footprint: The bubble enables circularly shift in vector B ($O(d)$) and alleviates the overhead of creating and fetching matrix ($O(d^2)$) of circularly shifted B as TPU-like systolic cell, reducing footprint by $d \times$. (2) Column-wise parallelism (CWP): The BS dataflow enables each column of a systolic cell to execute a circular convolution, thus multiple circular convolutions can be parallelized over multiple columns, which is not possible for GEMV in a TPU-like systolic cell. The low arithmetic intensity and memory-bound operations make GPUs inefficient for vector-symbolic circular convolution. For circular convolution of two d -dimensional vectors, the GPU arithmetic intensity is $d \times (d + d - 1) / (d \times d + d \times 1 + d \times 1)$, while CogSys arithmetic intensity is $d \times (d + d - 1) / (d \times 1 + d \times 1 + d \times 1)$. As in Fig. 11c, CogSys achieves peak performance when fully utilized, while GPU suffers from memory-bound. Additionally, GPUs require extra computations to handle the index calculations for the circularly shifted vector.

Cycle analysis of BS dataflow. Assuming the case of $nsPE$



	(a) Spatial Mapping	(b) Temporal Mapping
Latency	$k \times \lceil d / (N \times M) \rceil \times T$	$\lceil k / N \rceil \times \lceil d / M \rceil \times T$
Mem Reads under full util.	2d per T cycles	$(d + M) \times N$ per T cycles

Fig. 12: Adaptive spatial-temporal (ST) mapping. The ST mapping under BS dataflow. CogSys adaptively chooses the mapping scheme based on (N, M, k, d) workload and hardware configurations.

array size $M =$ input vector size d for vector-symbolic circular convolution, streaming the stationary input would take d cycles followed by input vectors taking $2d$ cycles to reach the final $nsPE$. Meanwhile, partial sums are aggregated along the array for the first output, followed by the remaining $(d - 1)$ outputs where each is generated per cycle. Thus, the end-to-end latency for vector-symbolic circular convolution of two d -dimensional vectors in a 1-D $nsPE$ compute array is $(4d - 1)$ cycles. In the case where $d \neq M$, latency T will be $(3M + d - 1)$ cycles where M for loading stationary vector, $2M$ for streaming vector reach final $nsPE$, and $(d - 1)$ cycles for remaining outputs.

D. Adaptive Spatial and Temporal Mapping Strategy

Spatial-temporal (ST) mapping flexibility. To efficiently support the various dimensions of vector-symbolic operations, we propose ST mapping featuring spatial mapping mode and temporal mapping mode (Fig. 12). The $nsPE$ array is easily expanded to N arrays. Spatial mapping, by parallelizing a single circular convolution into folds, reduces memory reads compared to temporal mapping, especially with many folds. Taking N arrays (M PEs each) as an example, spatial mapping requires $B_S = 2d$ memory reads per T cycles for d -dimensional vectors. while temporal mapping requires loading $B_T = (d + M) \times N$ elements per T cycles. Given neurosymbolic workloads typically have $d > 1000$, the bandwidth requirement (memory reads per T cycles) is reduced by $(N/2) \times$ via spatial

mapping. Temporal mapping, on the other hand, outperforms spatial mapping when $d < M$ by enabling the parallelization of multiple convolutions. For k vector-symbolic circular convolutions, temporal folding takes $C_T = \lceil k/N \rceil \times \lceil d/M \rceil \times T$ cycles, while spatial folding takes $C_S = k \times \lceil d/(N \times M) \rceil \times T$ cycles.

To efficiently process symbolic operations and balance between bandwidth and latency, we conduct an adaptive search between spatial and temporal mapping based on workloads and CogSys configurations. For example, For $N=32$ and $d=1024$ in NVSA ($k=210$) and LVRF ($k=2575$) workloads, CogSys opts for temporal mapping with 32 parallel circular convolutions.

E. Adaptive Scale-Up and Scale-Out Strategy

Scale-up and scale-out flexibility. To enhance design utilization and scalability, CogSys proposes to operate as a combination of scale-up and scaled-out reconfigurable arrays, with the support of systolic cell-wise parallelism (ScWP) and column-wise parallelism (CWP) for circular convolutions. The $(N=32, M=512)$ configuration is constructed from 16 32×32 cells by configuring the muxes to choose among five schemes, i.e., scale-up GEMM, scale-out GEMM, scale-up Conv, scale-out Conv, and scale-out GEMM+Conv, enabled by the systolic cell-wise heterogeneous partitioning scheme in Sec. VI-B. For GEMM, the scale-out scheme enables higher utilization and ScWP. For symbolic circular convolution, the scaled-out scheme enables ScWP and CWP for low-dimensional vectors.

Design space exploration. We search scale-out/scale-up schemes based on workloads and CogSys configurations to increase utilization and parallelism. For instance, the 16 32×32 scaled-out cells achieve 91.26% utilization, with $10.71 \times$ and $7.83 \times$ speedup over one 128×128 scaled-up and four 64×64 scaled-out cells, respectively, for NVSA and LVRF neural modules. For vector-symbolic operations, CogSys chooses a scale-up scheme for NVSA and LVRF (high-dimensional vector processing, $d=1024$) and a scale-out scheme for MIMONet (low-dimensional vector processing, $d=64$).

F. Double-buffered Memory and Custom SIMD Unit

Double-buffered memory. CogSys array is backed by three double-buffered SRAMs (Fig. 9b). The double-buffered memory is effective in reducing off-chip accesses and stalls due to loads and stores to reduce latency. SRAM A is common for all cells to utilize weights temporal reuse while SRAM B is distributed across cells. Through design space exploration, CogSys opts for 256kB for SRAM A and 4MB for SRAM B.

Custom SIMD units. CogSys employs a custom SIMD unit to execute vector reductions and element-wise operations (Fig. 9a). The SIMD unit efficiently handles data transfer between the CogSys array output and input SRAM, enabling the array to seamlessly access data for subsequent operations. The SIMD unit is comprised of multiple PEs, each designed with compact logic circuits (i.e., sum, mult/div, exp/log/tanh, norm, softmax, etc) to perform vector operations on quantized data. The adaptive workload-aware scheduling (Sec. VI) scheme schedules workloads across CogSys array and SIMD units to balance the runtime of neural and symbolic operations.

TABLE V: Design choice discussion. Area, latency, energy, and utilization comparison with reconfigurable and heterogeneous PEs.

	Configuration	Area	Latency	Energy	Utilization
Reconfigurable PE (CogSys)	$16 \times 32 \times 32$ reconfigurable neuro/symbolic PE	$1 \times$	$1 \times$	$1 \times$	90%
Heterogeneous PE	$16 \times 32 \times 32$ neuro PE $16 \times 32 \times 32$ symbolic PE	$1.96 \times$	$1 \times$	$1.3 \times$	45%
Heterogeneous PE	$8 \times 32 \times 32$ neuro PE $8 \times 32 \times 32$ symbolic PE	$0.98 \times$	$2 \times$	$1.3 \times$	45%

G. Design Choices Discussion

Reconfigurable or heterogeneous PE. While specialized PEs for neural and symbolic kernels may appear more efficient for simultaneous processing, our early-phase design space exploration reveals that using separate PEs for GEMM and circular convolution leads to hardware underutilization and increased area overhead due to the sequential execution of neural and symbolic kernels, as well as the varying proportions of these operations across different workloads. When comparing designs with the same chip size, specialized PEs result in longer latency, as they provide fewer effective compute units for either neural or symbolic tasks, as summarized in Tab. V. We thus opt for the reconfigurable PE approach, which offers lower area overhead and higher hardware utilization, making it suitable for both neuro-heavy and symbolic-heavy workloads.

VI. COGSYS: SCHEDULING STRATEGY

This section presents CogSys adaptive workload-aware scheduling strategy. We first identify the system-level challenges of neurosymbolic workloads (Sec. VI-A), and then introduce CogSys scheduling scheme (Sec. VI-B) and further discuss its scalability to other workloads and tasks (Sec. VI-C).

A. Neurosymbolic System-Level Challenges

We identify two main neurosymbolic system-level challenges (Fig. 13a): **First**, the sequential execution and frequent interactions of neural and symbolic components results in long latency and low system throughput. **Second**, the heterogeneous neural and symbolic kernels result in low compute array utilization and efficiency of ML accelerator.

B. Adaptive Workload-Aware Scheduling (adSCH) Strategy

Adaptive scheduling (adSCH) strategy. To solve the system-level challenges, CogSys features an *adSCH* scheme and greatly improves hardware utilization and performance. (1) **Interleaved neural/symbolic processing.** Despite the dependencies in neural and symbolic tasks, symbolic operations of other tasks can be interleaved within neural layer of current task via reconfigurable neuro/symbolic PE arrays (Fig. 13b). (2) **Adaptive neuro/symbolic array partition strategy.** We propose to adaptively allocate CogSys cells to various neural and symbolic kernels (cell-wise partition), and allocate symbolic cell columns to parallel circular convolution operations (column-wise partition) (Fig. 13c). This partition strategy is effective in handling both neural- and symbolic-intensive workloads and promotes parallelism and hardware utilization.

Scheduling Implementation. CogSys workload-aware scheduling is performed offline by software. Since the model architecture, size, and data are known prior to execution,

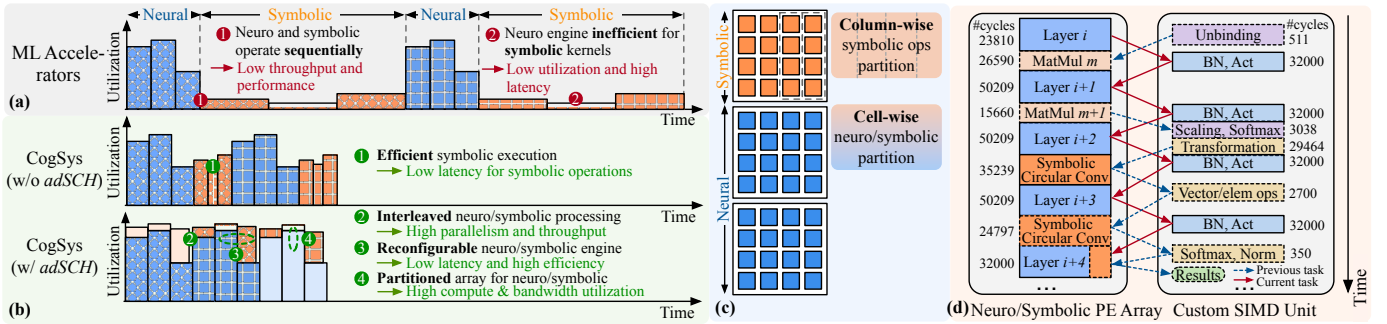


Fig. 13: Adaptive workload-aware scheduling (*adSCH*) strategy. (a) Neurosymbolic system-level challenges. The *adSCH* scheme enables (b) interleaved neural/symbolic processing and (c) cell-wise partition across CogSys arrays with multi-level parallelism. (d) An *adSCH* example on NVSA algorithm. The scheduling method ensures CogSys is adaptable and scalable across neurosymbolic workloads and tasks.

TABLE VI: Baseline. The specifications of hardware baseline.

HW	General Purpose Processor/SoC					CogSys (Ours)	HW	ML Accelerator			CogSys (Ours)
	CPU Xeon	RTX GPU	TX2	NX				TPU-like	MTIA-like	Gemmini-like	
Power	145W	250W	15W	20W	1.48W	SRAM	4.5MB	4.5MB	4.5MB	4.5MB	
						#PE	1 128×128	16 32×32	64 16×16	16 32×32	

the host CPU precomputes the mapping of operations and array configurations, which are then offloaded to CogSys. This ensures optimal or near-optimal scheduling with zero runtime latency. The scheduling process uses a greedy search algorithm: (1) Generate an operation graph based on operation type, size, dependencies, and number of iterations; (2) Assign ready operations (not blocked by dependencies) to newly available cells, with runtime estimated analytically; (3) Maximize utilization by prioritizing neural tasks for larger cell blocks and symbolic tasks for smaller ones. Since the search only considers available blocks within the 16 array cells and ready tasks, the search space is limited to $<O(10^3)$ per time step, resulting in minimal offline overhead and no runtime overhead.

Adaptive scheduling example. Fig. 13d presents a detailed example of *adSCH* scheme with operations and cycle numbers in a NVSA segment [33]. CogSys reconfigurable array schedules neural (convolutions, GEMMs) and symbolic (circular convolutions), while element-wise operations are offloaded to SIMD units. To mitigate underutilization, CogSys executes VSA-based codebook and symbolic kernels of the previous batch on idle hardware pieces during neural layers of the current batch, thus eliminating symbolic bottleneck. Particularly, *multi-level parallelism* is adopted to process different symbolic rules and attributes to further improve efficiency.

C. Scalability and Variability Support

Scalable across neurosymbolic workloads and cognition tasks. The *adSCH* technique enables CogSys to be easily reconfigured across (1) neurosymbolic workloads (e.g., NVSA, MIMONets, LVRf, etc) and (2) cognitive tasks such as procedurally generated matrices (PGM) [11], compositional visual reasoning (CVR) [94], synthetic visual reasoning test (SVRT) [20] with different attributes and rules (Fig. 2, Tab. I). Coupled with *nsPE* reconfigurable arrays, *BS* dataflow, and *ST* mapping, *adSCH* scheme ensures symbolic operations interleaved with neural operations with high throughput, enabling various kinds of VSA-based neurosymbolic workloads to be executed on CogSys with high efficiency and utilization,

and adapt to different neuro-symbolic workload ratios and unpredictably changing workloads.

VII. EVALUATION RESULTS

This section first introduces the detailed settings for evaluating our proposed CogSys framework (Sec. VII-A), and then benchmarks our proposed CogSys algorithm optimization (Sec. VII-B) and accelerator (Sec. VII-C), demonstrating the practical of efficient and scalable neurosymbolic systems.

A. Experimental Setup

Datasets. To evaluate the achieved cognitive reasoning capability of CogSys, we conduct experiments on the commonly-used spatial-temporal reasoning RAVEN [95], I-RAVEN [36], PGM [11], CVR [94], and SVRT [20]. The task performance is measured by the probabilistic abduction accuracy.

Algorithm setup. We evaluate CogSys on three state-of-the-art VSA-based neurosymbolic workloads, i.e., NVSA [33], MIMONet [60], and LVRf [32]. Following [32], [33], [60], we determine the training hyperparameters based on the end-to-end reasoning performance on the validation set.

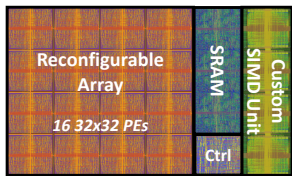
Baselines. We consider several hardware baselines, including TX2, Xavier NX, RTX GPU, Xeon CPU, and ML accelerators (TPU, MTIA, Gemmini). Tab. VI lists their configurations.

Hardware setup. To evaluate energy and area of CogSys accelerator, we implement CogSys in RTL, synthesize using Synopsys Design Compiler [79] with 0.8 GHz, and place and route using Cadence Innovus [13] based on TSMC 28nm node. Fig. 14 illustrates the layout and specifications of CogSys accelerator. In addition, we develop a cycle-accurate simulator to estimate CogSys accelerator performance on different reasoning tasks. The proposed CogSys accelerator consumes an area of 4.0 mm² and an average power consumption of 1.48 W. *Compared with conventional systolic arrays that only support neural operations, CogSys provides reconfigurable support for neural and symbolic operations with only 4.8% area overhead.*

B. CogSys Algorithm Optimization Performance

Factorization accuracy. To assess the effectiveness of our factorization and stochasticity methods, we compare CogSys with the state-of-the-art factorizer [50] across 14 test cases (Tab. VII). The results show a slight improvement in factorization accuracy for object constituent attribute extraction.

Layout of Proposed CogSys Accelerator



Accelerator Specs

Technology	28 nm	DRAM BW	700 GB/s
#Arrays	16	Frequency	800 MHz
Size of Each Array	32x32	Voltage	1 V
#SIMD PEs	512	Power	1.48 W
SRAM	4.5 MB	Area	4.0 mm ²

Fig. 14: CogSys accelerator. The layout and performance specifications of our proposed CogSys accelerator.

TABLE VII: Factorization accuracy comparison. Factorization accuracy for object constituent attribute estimation across 14 scenarios.

Test	2×2 Grid	3×3 Grid	Left-Right	Up-Down	Center	O-IC	DistFour	Average
[50]	95.8%	94.7%	96.1%	95.6%	94.9%	95.3%	94.5%	95.3%
CogSys	95.7%	95.2%	96.1%	95.7%	95.3%	95.5%	94.4%	95.4%
Test	Constant	Progression	XOR	AND	OR	Arithmetic	Distribution	Average
[50]	93.3%	93.5%	93.9%	93.7%	93.5%	93.1%	92.7%	93.4%
CogSys	93.3%	93.6%	93.9%	93.6%	93.7%	93.4%	92.7%	93.5%

TABLE VIII: CogSys algorithm optimization performance. Compared with NVSA, CogSys exhibits comparable reasoning capability with smaller memory footprint requirement, achieved through the proposed factorization, stochasticity, and quantization techniques.

Datasets	NVSA [33]	CogSys (+Factorization & Stoch.)	CogSys (+Quant.)
RAVEN [95]	98.5%	98.7±0.3%	98.6±0.4%
I-RAVEN [36]	99.0%	99.0±0.3%	98.8±0.4%
PGM [11]	68.3%	68.6±0.8%	68.4±1.0%
#Parameters	38 MB	32 MB	8 MB

TABLE IX: Efficiency improvement from optimized precision. CogSys optimizes NVSA algorithm to INT8 to enable hardware area and power savings while maintaining the reasoning capability.

Arithmetic Precision		FP32	FP8	INT8
CogSys Accuracy (NVSA=98.5%)		98.9%	98.9%	98.7%
Reconfigurable Array 16 32×32 PEs	Area (mm ²)	28.9	9.9	3.8
	Power (mW)	4468.5	1237.8	1104.6
Custom SIMD Unit 512 PEs	Area (mm ²)	2.01	0.28	0.21
	Power (mW)	297.0	64.8	80.4
Reconfig. Array Area Overhead vs. SA		<1%	4.8%	12.1%

Reasoning accuracy. To evaluate CogSys algorithm optimization (Sec. IV), we benchmark it on five reasoning tasks in terms of the achieved accuracy (Sec. VII-A). Tab. VIII uses NVSA as an example and benchmarks on RAVEN, I-RAVEN, and PGM datasets, we observe that CogSys achieves comparable reasoning accuracy through factorization and injected stochasticity. Through quantization, CogSys enables 4.75× memory footprint savings as well as 7.71× area and 4.02× power savings (Tab. IX) under TSMC 28nm technology node. We get consistent observations in MIMONet and LVRf workloads under CVR and SVRT datasets as well.

C. CogSys Accelerator Performance

Performance improvement. We benchmark CogSys accelerator with RTX GPU, Xeon CPU, and edge SoCs (Jetson TX2, Xavier NX) for accelerating neurosymbolic algorithms on five reasoning tasks (Fig. 15) with different difficulty levels. For GPU baseline, for neuro kernels, we use Pytorch package that leverages CUDA and cuBLAS/cuDNN libraries; for symbolic kernels, we implement custom kernels optimized for vector-symbolic operations. The workload is tiled by CuDNN

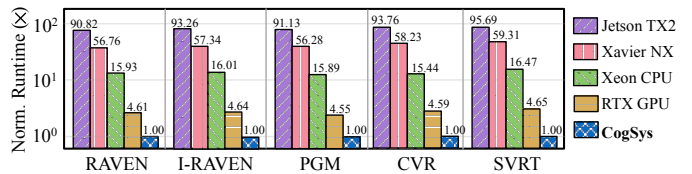


Fig. 15: End-to-end runtime improvement. CogSys consistently outperforms Xeon CPU, RTX GPU, and edge SoCs (TX2, NX) in end-to-end runtime evaluated on five spatial-temporal reasoning tasks.

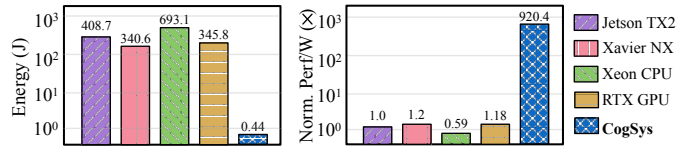


Fig. 16: Energy efficiency improvement. CogSys consistently reduces energy consumption and improves performance per watt compared to CPU and GPUs, evaluated from five reasoning tasks.

in Pytorch based on block sizes that fit well in GPU memory. We observe that CogSys exhibits consistent speedup across datasets, e.g., 90.82×/56.76× speedup over TX2 and NX, indicating its high efficiency and scalability capability. Furthermore, CogSys achieves real-time performance (<0.3 s) [33] for solving logical reasoning tasks, indicating that CogSys is the *first* to enable real-time neurosymbolic system with superior reasoning and generalization capability, offering a promising solution for future cognitive applications.

Energy efficiency improvement. We benchmark CogSys accelerator on energy consumption and efficiency on five reasoning tasks (Fig. 16). We can observe that CogSys accelerator achieves two orders of energy efficiency than RTX GPU, Xeon CPU, TX2, and NX, indicating its efficiency and applicability to resource-constrained neurosymbolic systems. To further assess CogSys energy efficiency in long-term deployment, we conduct consecutive tests on CogSys using mixed workloads, incorporating both high-demand and low-activity periods, with 10-second idle intervals between scenarios. On average, CogSys achieves 730× energy efficiency compared to RTX GPU. Additionally, when compared to V100 and A100 GPUs, CogSys shows 4.43× and 1.43× speedup, with 748× and 241× energy efficiency, respectively.

Comparison with TPU/GPU. We benchmark symbolic circular convolution over TPU-like SA (with the same number of PEs) and GPU under different vector dimensions and number of operations (Fig. 17). We observe CogSys reconfigurable array achieves up to 75.96× and 18.90× speedup over TPU-like SA and GPU, and is effective in both low-dimension and high-dimension vector-symbolic operations.

Comparison with ML accelerators. We benchmark the runtime of neural and symbolic operations on TPU [41], Gemini [28], and MTIA [19]-like architecture over different neurosymbolic models and tasks (Fig. 18). For a fair comparison, we keep all hardware configurations with the same number of PEs. Compared with current ML accelerators, we observe that CogSys achieves similar performance in neural operations, while exhibiting superior symbolic operation efficiency thus end-to-end speedup in neurosymbolic systems. Additionally,

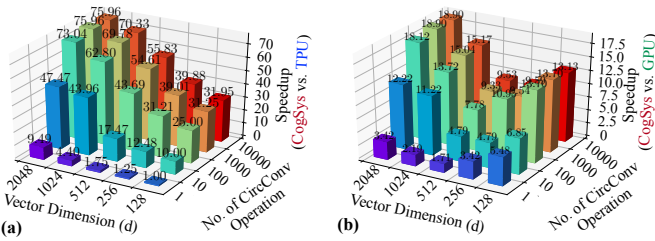


Fig. 17: Improved efficiency over TPU/GPU. Speedup comparison of circular convolution on CogSys, TPU-like systolic array and GPU, CogSys shows up to 75.96 \times and 18.90 \times runtime improvement.

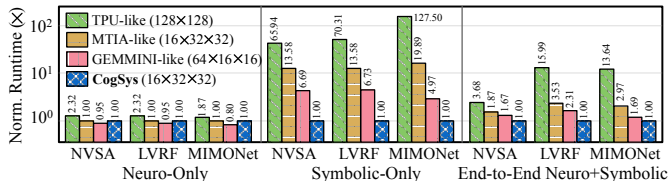


Fig. 18: Improved efficiency over ML accelerators. Speedup comparison of neural, symbolic, and end-to-end neurosymbolic system over TPU [41], Gemmini [28], and MTIA [19]-like architecture.

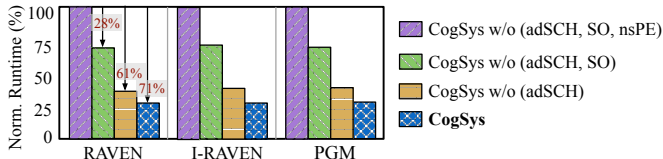


Fig. 19: Ablation Study on CogSys Accelerator Techniques. The runtime achieved by CogSys w/o the adaptive scheduling (adSCH), scalable array (SO), and reconfigurable PE (nsPE) across tasks.

we compare CogSys with hyperdimensional computing accelerator [37] across models and tasks and observe a 7.2 \times average speedup. This improvement is mainly due to the lack of efficient neuro and symbolic support and circular convolution handling in hyperdimensional computing architectures.

Ablation study on the proposed hardware techniques.

As illustrated in Sec. V and Sec. VI, CogSys features reconfigurable neuro/symbolic PE with bubble streaming dataflow and spatial-temporal mapping, scalable array architecture, and adaptive scheduling strategy to reduce compute latency and memory footprint for neural and symbolic kernels. To verify the effectiveness of our proposed methods, we summarize the runtime of CogSys w/o the scheduling, scalable architecture, and reconfigurable PE in Fig 19. In particular, the proposed scheduling strategy can trim down the runtime by 28% on average. Additionally, with the proposed scalable array and reconfigurable PE, the runtime reduction ratio can be further enlarged to 61% and 71%, indicating that both proposed techniques are necessary for our CogSys accelerator to achieve the desired efficient and scalable reasoning capability.

Ablation study of necessity of co-design. To the best of our knowledge, our proposed CogSys, as an algorithm-hardware co-design framework, is the first that has achieved efficient and scalable on-device neurosymbolic-based system. To verify the necessity of such co-design strategy, we summarize the runtime of our CogSys w/o the proposed algorithm optimiza-

TABLE X: Ablation study of necessity of co-design. The normalized runtime achieved by CogSys framework w/o the proposed algorithm optimization or hardware techniques on different tasks.

Neurosymbolic Cognitive Solution Algorithm @ Hardware	Normalized Runtime (%) on				
	RAVEN [95]	I-RAVEN [36]	PGM [11]	CVR [94]	SVRT [20]
NVSA [33] @ Xavier NX	100	100	100	100	100
CogSys Algorithm @ Xavier NX	89.5%	88.9%	90.7%	87.6%	88.4%
CogSys Algorithm @ CogSys Accelerator	1.76%	1.74%	1.78%	1.72%	1.69%

tion or hardware techniques in Tab. X. Specifically, with our proposed CogSys algorithm optimization, we can trim down the runtime to 89.5% as compared to NVSA [33] on the same Xavier NX hardware and RAVEN task. Moreover, with both proposed CogSys algorithm optimization and accelerator, the runtime can be reduced to 1.76%, indicating the necessity of the co-design strategy of CogSys framework.

VIII. RELATED WORK

Neurosymbolic AI. Neurosymbolic AI holds significant potential for enhancing trustworthiness, reasoning, and robustness of next-generation cognitive applications where agents can make decisions in an explainable manner, and intelligence is pervasively embedded in human-AI interactions [10], [18], [33], [34], [56], [66], [92], [93], [97]. Current neurosymbolic research mostly focuses on algorithms; however, the lack of attention to its inefficiency on off-the-shelf hardware may hinder neurosymbolic AI development in the long run. *CogSys thus takes the first step to understand neurosymbolic architectural and system characteristics and proposes a co-design framework to make it more efficient and deployable at scale.*

Accelerators for emerging applications. With the slowdown of technology scaling, custom architecture is a pragmatic approach to ensure simultaneous improvements in performance and efficiency. Beyond DNNs [21], [40], [69], [71], [77], [80], [81], [99], hardware acceleration has been found effective in emerging applications such as genome sequencing [22], [23], graph [25], [75], mobile vision [54], [55], [85], drone [14], [47], [48], robotics [30], [51], [52], [58], [63], privacy and security [27], [62], [73], [74], etc. *Despite the presence of these accelerators, CogSys is the first proposal to offer reconfigurable support for both neural and symbolic kernels, facilitating efficient and scalable neurosymbolic systems.*

IX. CONCLUSION

To enable efficient and scalable neurosymbolic AI towards real-time cognitive applications, we propose CogSys, the first algorithm-hardware co-design framework dedicated to accelerating neurosymbolic AI. CogSys identifies the unique opportunities for neurosymbolic acceleration, including efficient factorization, reconfigurable neural/symbolic PE, bubble streaming dataflow, and adaptive scheduler, leveraging which we develop algorithm optimizations and dedicated accelerator. We believe CogSys can open up an exciting perspective toward efficient and scalable cognitive reasoning systems at scale.

ACKNOWLEDGEMENTS

This work was supported in part by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

- [1] “Neurips workshop on neuro causal and symbolic ai (ncsi),” 2022. [Online]. Available: <https://neurips.cc/virtual/2022/workshop/50011>
- [2] “Aaai tutorial on advances in neuro symbolic reasoning and learning,” 2023. [Online]. Available: <https://neurosymbolic.asu.edu/2023-aaai-tutorial-advances-in-neuro-symbolic-reasoning/>
- [3] “Ibm neuro-symbolic ai workshop,” 2023. [Online]. Available: <https://ibm.github.io/neuro-symbolic-ai/>
- [4] “Neuro-symbolic ai summer school,” 2023. [Online]. Available: <https://neurosymbolic.github.io/nsss2023/>
- [5] “1st international conference on neuro-symbolic systems (neus),” May 2024. [Online]. Available: <https://www.neusconference.org>
- [6] “Aaai workshop on neuro-symbolic learning and reasoning in the era of large language models (nuclear),” 2024. [Online]. Available: <https://nuclear-workshop.github.io/aaai2024/>
- [7] “Ijcai first international workshop on logical foundations of neuro-symbolic ai (Insai 2024),” 2024. [Online]. Available: <https://sites.google.com/view/Insai2024/>
- [8] “International conference on neuro-symbolic learning and reasoning (nesy),” 2024. [Online]. Available: <https://sites.google.com/view/nesy2024>
- [9] A. G. Anderson, K. Ratnam, A. Roorda, and B. A. Olshausen, “High-acuity vision from retinal image motion,” *Journal of vision*, vol. 20, no. 7, pp. 34–34, 2020.
- [10] S. Badreddine, A. d. Garcez, L. Serafini, and M. Spranger, “Logic tensor networks,” *Artificial Intelligence*, vol. 303, p. 103649, 2022.
- [11] D. Barrett, F. Hill, A. Santoro, A. Morcos, and T. Lillicrap, “Measuring abstract reasoning in neural networks,” in *International conference on machine learning (ICML)*. PMLR, 2018, pp. 511–520.
- [12] G. Booch, F. Fabiano, L. Horesh, K. Kate, J. Lenchner, N. Linck, A. Loreggia, K. Murgesan, N. Mattei, F. Rossi *et al.*, “Thinking fast and slow in ai,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15 042–15 046.
- [13] Cadence, “Innovus implementation system - cadence,” https://www.cadence.com/en_US/home/tools/digital-design-and-signoff/soc-implementation-and-floorplanning/innovus-implementation-system.html.
- [14] M. Chang, A. S. Lele, S. D. Spetalnick, B. Crafton, S. Konno, Z. Wan, A. Bhat, W.-S. Khwa, Y.-D. Chih, M.-F. Chang *et al.*, “A 73.53 tops/w 14.74 tops heterogeneous rram in-memory and sram near-memory soc for hybrid frame and event-based target tracking,” in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 426–428.
- [15] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, “Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks,” *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 52, no. 1, pp. 127–138, 2016.
- [16] E. Chung, J. Fowers, K. Ovtcharov, M. Papamichael, A. Caulfield, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman *et al.*, “Serving dnns in real time at datacenter scale with project brainwave,” *IEEE Micro*, vol. 38, no. 2, pp. 8–20, 2018.
- [17] K. Daniel, *Thinking, fast and slow*, 2017.
- [18] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, and D. Zhou, “Neural logic machines,” in *International Conference on Learning Representations (ICLR)*, 2019.
- [19] A. Firoozshahian, J. Coburn, R. Levenstein, R. Nattoji, A. Kamath, O. Wu, G. Grewal, H. Aepala, B. Jakka, B. Dreyer *et al.*, “Mtia: First generation silicon targeting meta’s recommendation systems,” in *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA)*, 2023, pp. 1–13.
- [20] F. Fleuret, T. Li, C. Dubout, E. K. Wampler, S. Yantis, and D. Geman, “Comparing machines and humans on a visual categorization test,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 43, pp. 17 621–17 625, 2011.
- [21] J. Fowers, K. Ovtcharov, M. Papamichael, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, L. Adams, M. Ghandi *et al.*, “A configurable cloud-scale dnn processor for real-time ai,” in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 1–14.
- [22] D. Fujiki, A. Subramanian, T. Zhang, Y. Zeng, R. Das, D. Blaauw, and S. Narayanasamy, “Genax: A genome sequencing accelerator,” in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 69–82.
- [23] D. Fujiki, S. Wu, N. Ozog, K. Goliya, D. Blaauw, S. Narayanasamy, and R. Das, “Seedex: A genome sequencing accelerator for optimal alignments in subminimal space,” in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 937–950.
- [24] P. M. Furlong and C. Eliasmith, “Modelling neural probabilistic computation using vector symbolic architectures,” *Cognitive Neurodynamics*, pp. 1–24, 2023.
- [25] C. Gao, M. Afarin, S. Rahman, N. Abu-Ghazaleh, and R. Gupta, “Mega evolving graph accelerator,” in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2023, pp. 310–323.
- [26] A. d. Garcez and L. C. Lamb, “Neurosymbolic ai: The 3 rd wave,” *Artificial Intelligence Review*, pp. 1–20, 2023.
- [27] R. Geelen, M. Van Beirendonck, H. V. Pereira, B. Huffman, T. McAuley, B. Selfridge, D. Wagner, G. Dimou, I. Verbauwhede, F. Vercauteren *et al.*, “Basalisc: Programmable asynchronous hardware accelerator for bgv fully homomorphic encryption,” *Cryptology ePrint Archive*, 2022.
- [28] H. Genc, S. Kim, A. Amid, A. Haj-Ali, V. Iyer, P. Prakash, J. Zhao, D. Grubb, H. Liew, H. Mao *et al.*, “Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration,” in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 769–774.
- [29] C. Han, J. Mao, C. Gan, J. Tenenbaum, and J. Wu, “Visual concept-metaconcept learning,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [30] Y. Hao, Y. Gan, B. Yu, Q. Liu, Y. Han, Z. Wan, and S. Liu, “Orianna: An accelerator generation framework for optimization-based robotic applications,” in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Volume 2*, 2024, pp. 813–829.
- [31] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, “Interpreting black-box models: a review on explainable artificial intelligence,” *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.
- [32] M. Hersche, F. Di Stefano, T. Hofmann, A. Sebastian, and A. Rahimi, “Probabilistic abduction for visual abstract reasoning via learning rules in vector-symbolic architectures,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [33] M. Hersche, M. Zeqiri, L. Benini, A. Sebastian, and A. Rahimi, “A neuro-vector-symbolic architecture for solving raven’s progressive matrices,” *Nature Machine Intelligence*, vol. 5, no. 4, pp. 363–375, 2023.
- [34] P. Hohenecker and T. Lukas, “Ontology reasoning with deep neural networks,” *Journal of Artificial Intelligence Research*, vol. 68, pp. 503–540, 2020.
- [35] J. Hsu, J. Mao, and J. Wu, “Ns3d: Neuro-symbolic grounding of 3d objects and relations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2614–2623.
- [36] S. Hu, Y. Ma, X. Liu, Y. Wei, and S. Bai, “Stratified rule-aware network for abstract visual reasoning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1567–1574.
- [37] M. Ibrahim, Y. Kim, and J. M. Rabaey, “Efficient design of a hyperdimensional processing unit for multi-layer cognition,” in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2024, pp. 1–6.
- [38] M. Ibrahim, Z. Wan, H. Li, P. Panda, T. Krishna, P. Kanerva, Y. Chen, and A. Raychowdhury, “Special session: Neuro-symbolic architecture meets large language models: A memory-centric perspective,” in *2024 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)*. IEEE, 2024, pp. 11–20.
- [39] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [40] N. Jouppi, G. Kurian, S. Li, P. Ma, R. Nagarajan, L. Nai, N. Patil, S. Subramanian, A. Swing, B. Towles *et al.*, “Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,” in *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA)*, 2023, pp. 1–14.
- [41] N. P. Jouppi, D. H. Yoon, M. Ashcraft, M. Gottscho, T. B. Jablin, G. Kurian, J. Laudon, S. Li, P. Ma, X. Ma *et al.*, “Ten lessons from three generations shaped google’s tpuv4i: Industrial product,” in

- 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2021, pp. 1–14.
- [42] N. P. Jouppi, D. H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, and D. Patterson, “A domain-specific supercomputer for training deep neural networks,” *Communications of the ACM*, vol. 63, no. 7, pp. 67–78, 2020.
- [43] A. Kalyanpur, K. Saravanakumar, V. Barres, J. Chu-Carroll, D. Melville, and D. Ferrucci, “Llm-arc: Enhancing llms with an automated reasoning critic,” *arXiv preprint arXiv:2406.17663*, 2024.
- [44] M. Kang and B. Li, “R²-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning,” *arXiv preprint arXiv:2407.05557*, 2024.
- [45] D. Kim, J. Kung, S. Chai, S. Yalamanchili, and S. Mukhopadhyay, “Neurocube: A programmable digital neuromorphic architecture with high-density 3d memory,” *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 380–392, 2016.
- [46] D. Kleyko, M. Davies, E. P. Frady, P. Kanerva, S. J. Kent, B. A. Olshausen, E. Osipov, J. M. Rabaey, D. A. Rachkovskij, A. Rahimi *et al.*, “Vector symbolic architectures as a computing framework for emerging hardware,” *Proceedings of the IEEE*, vol. 110, no. 10, pp. 1538–1571, 2022.
- [47] S. Krishnan, Z. Wan, K. Bhardwaj, N. Jadhav, A. Faust, and V. J. Reddi, “Roofline model for uavs: A bottleneck analysis tool for onboard compute characterization of autonomous unmanned aerial vehicles,” in *2022 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2022, pp. 162–174.
- [48] S. Krishnan, Z. Wan, K. Bhardwaj, P. Whatmough, A. Faust, S. Neuman, G.-Y. Wei, D. Brooks, and V. J. Reddi, “Automatic domain-specific soc design for autonomous unmanned aerial vehicles,” in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022, pp. 300–317.
- [49] J. Kwon, J. Tenenbaum, and S. Levine, “Neuro-symbolic models of human moral judgment,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.
- [50] J. Langenegger, G. Karunaratne, M. Hersche, L. Benini, A. Sebastian, and A. Rahimi, “In-memory factorization of holographic perceptual representations,” *Nature Nanotechnology*, vol. 18, no. 5, pp. 479–485, 2023.
- [51] Q. Liu, Z. Wan, B. Yu, W. Liu, S. Liu, and A. Raychowdhury, “An energy-efficient and runtime-reconfigurable fpga-based accelerator for robotic localization systems,” in *2022 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2022, pp. 01–02.
- [52] S. Liu, Z. Wan, B. Yu, and Y. Wang, *Robotic computing on fpgas*. Springer, 2021, vol. 16.
- [53] Y. Liu, R. Ryskin, R. Futral, and E. Gibson, “A verb-frame frequency account of constraints on long-distance dependencies in english,” *Cognition*, vol. 222, p. 104902, 2022.
- [54] Z.-G. Liu, P. N. Whatmough, and M. Mattina, “Systolic tensor array: An efficient structured-sparse gemm accelerator for mobile cnn inference,” *IEEE Computer Architecture Letters*, vol. 19, no. 1, pp. 34–37, 2020.
- [55] Z.-G. Liu, P. N. Whatmough, Y. Zhu, and M. Mattina, “S2ta: Exploiting structured sparsity for energy-efficient mobile cnn acceleration,” in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2022, pp. 573–586.
- [56] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, and L. De Raedt, “Neural probabilistic logic programming in deepproblog,” *Artificial Intelligence*, vol. 298, p. 103504, 2021.
- [57] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision,” *International Conference on Learning Representations (ICLR)*, 2019.
- [58] V. Mayoral-Vilches, J. Jabbour, Y.-S. Hsiao, Z. Wan, M. Crespo-Álvarez, M. Stewart, J. M. Reina-Muñoz, P. Nagras, G. Vikhe, M. Bakhshalipour *et al.*, “Robotperf: An open-source, vendor-agnostic, benchmarking suite for evaluating robotics computing system performance,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 8288–8297.
- [59] L. Mei, J. Mao, Z. Wang, C. Gan, and J. B. Tenenbaum, “Falcon: fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations,” *International Conference on Learning Representations (ICLR)*, 2022.
- [60] N. Menet, M. Hersche, G. Karunaratne, L. Benini, A. Sebastian, and A. Rahimi, “Mimonets: Multiple-input-multiple-output neural networks exploiting computation in superposition,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
- [61] T. Mu, A. Helyar, J. Heidecke, J. Achiam, A. Vallone, I. Kivlichen, M. Lin, A. Beutel, J. Schulman, and L. Weng, “Rule based rewards for language model safety,” *Open AI*, 2024.
- [62] M. Nabeel, D. Soni, M. Ashraf, M. A. Gebremichael, H. Gamil, E. Chielle, R. Karri, M. Sanduleanu, and M. Maniatakos, “Cofhee: A co-processor for fully homomorphic encryption execution,” in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–2.
- [63] S. M. Neuman, R. Ghosal, T. Bourgeat, B. Plancher, and V. J. Reddi, “Roboshape: Using topology patterns to scalably and flexibly deploy accelerators across robots,” in *Proceedings of the 50th Annual International Symposium on Computer Architecture (ISCA)*, 2023, pp. 1–13.
- [64] C. Núñez-Molina, P. Mesejo, and J. Fernández-Olivares, “A review of symbolic, subsymbolic and hybrid methods for sequential decision making,” *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–36, 2024.
- [65] L. I. G. Olascoaga, A. Menon, M. Ibrahim, and J. Rabaey, “A brain-inspired hierarchical reasoning framework for cognition-augmented prosthetic grasping,” in *Combining Learning and Reasoning: Programming Languages, Formalisms, and Representations*, 2021.
- [66] C. Pryor, C. Dickens, E. Augustine, A. Albalak, W. Wang, and L. Getoor, “Neupsl: Neural probabilistic soft logic,” *arXiv preprint arXiv:2205.14268*, 2022.
- [67] D. A. Rachkovskij and E. M. Kussul, “Binding and normalization of binary sparse distributed representations by context-dependent thinning,” *Neural Computation*, vol. 13, no. 2, pp. 411–452, 2001.
- [68] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning (ICML)*. PMLR, 2021, pp. 8748–8763.
- [69] A. Ramachandran, Z. Wan, G. Jeong, J. Gustafson, and T. Krishna, “Algorithm-hardware co-design of distribution-aware logarithmic-posit encodings for efficient dnn inference,” in *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC)*, 2024, pp. 1–6.
- [70] B. Romera-Paredes, M. Barekatin, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi *et al.*, “Mathematical discoveries from program search with large language models,” *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.
- [71] A. Samajdar, P. Mannan, K. Garg, and T. Krishna, “Genesys: Enabling continuous learning through neural network evolution in hardware,” in *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2018, pp. 855–866.
- [72] A. Samajdar, E. Qin, M. Pellauer, and T. Krishna, “Self adaptive reconfigurable arrays (sara) learning flexible gemm accelerator configuration and mapping-space using ml,” in *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, 2022, pp. 583–588.
- [73] N. Samardzic, A. Feldmann, A. Krastev, S. Devadas, R. Dreslinski, C. Peikert, and D. Sanchez, “F1: A fast and programmable accelerator for fully homomorphic encryption,” in *54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2021, pp. 238–252.
- [74] N. Samardzic, A. Feldmann, A. Krastev, N. Manohar, N. Genise, S. Devadas, K. Eldefrawy, C. Peikert, and D. Sanchez, “Craterlake: a hardware accelerator for efficient unbounded computation on encrypted data,” in *Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA)*, 2022, pp. 173–187.
- [75] N. Shah, W. Meert, and M. Verhelst, “Dpu-v2: Energy-efficient execution of irregular directed acyclic graphs,” in *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2022, pp. 1288–1307.
- [76] V. Shah, A. Sharma, G. Shroff, L. Vig, T. Dash, and A. Srinivasan, “Knowledge-based analogical reasoning in neuro-symbolic latent spaces,” *arXiv preprint arXiv:2209.08750*, 2022.
- [77] Y. S. Shao, J. Clemons, R. Venkatesan, B. Zimmer, M. Fojtik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina *et al.*, “Simba: Scaling deep-learning inference with multi-chip-module-based architecture,” in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2019, pp. 14–27.
- [78] A. Sheth and K. Roy, “Neurosymbolic value-inspired artificial intelligence (why, what, and how),” *IEEE Intelligent Systems*, vol. 39, no. 1, pp. 5–11, 2024.

- [79] Synopsys, "Design compiler - synopsys," <https://www.synopsys.com/implementation-and-signoff/rtl-synthesis-test/dc-ultra.html>.
- [80] T. Tambe, E.-Y. Yang, G. G. Ko, Y. Chai, C. Hooper, M. Donato, P. N. Whatmough, A. M. Rush, D. Brooks, and G.-Y. Wei, "A 16-nm soc for noise-robust speech and nlp edge ai inference with bayesian sound source separation and attention-based dnns," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 58, no. 2, pp. 569–581, 2022.
- [81] T. Tambe, E.-Y. Yang, Z. Wan, Y. Deng, V. J. Reddi, A. Rush, D. Brooks, and G.-Y. Wei, "Algorithm-hardware co-design of adaptive floating-point encodings for resilient deep learning inference," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2020, pp. 1–6.
- [82] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," *International Conference on Learning Representations (ICLR)*, 2021.
- [83] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, no. 7995, pp. 476–482, 2024.
- [84] Z. Wan, A. Anwar, Y.-S. Hsiao, T. Jia, V. J. Reddi, and A. Raychowdhury, "Analyzing and improving fault tolerance of learning-based navigation systems," in *2021 58th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2021, pp. 841–846.
- [85] Z. Wan, C.-K. Liu, M. Ibrahim, H. Yang, S. Spetalnick, T. Krishna, and A. Raychowdhury, "H3dfact: Heterogeneous 3d integrated cim for factorization with holographic perceptual representations," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2024, pp. 1–6.
- [86] Z. Wan, C.-K. Liu, H. Yang, R. Raj, C. Li, H. You, Y. Fu, C. Wan, S. Li, Y. Kim *et al.*, "Towards efficient neuro-symbolic ai: From workload characterization to hardware architecture," *IEEE Transactions on Circuits and Systems for Artificial Intelligence (TCASAI)*, 2024.
- [87] Z. Wan, C.-K. Liu, H. Yang, R. Raj, C. Li, H. You, Y. Fu, C. Wan, A. Samajdar, Y. C. Lin *et al.*, "Towards cognitive ai systems: Workload and characterization of neuro-symbolic ai," in *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2024, pp. 268–279.
- [88] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for vision and vision-language tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 175–19 186.
- [89] Y. Wang, Y. Han, C. Wang, S. Song, Q. Tian, and G. Huang, "Computation-efficient deep learning for computer vision: A survey," *Cybernetics and Intelligence*, 2024.
- [90] A. Xiao, J. Huang, D. Guan, X. Zhang, S. Lu, and L. Shao, "Unsupervised point cloud representation learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.
- [91] H. Xiong, Z. Wang, X. Li, J. Bian, Z. Xie, S. Mumtaz, and L. E. Barnes, "Converging paradigms: The synergy of symbolic and connectionist ai in llm-empowered autonomous agents," *arXiv preprint arXiv:2407.08516*, 2024.
- [92] Z. Yang, A. Ishay, and J. Lee, "Neurasp: Embracing neural networks into answer set programming," in *29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, 2020.
- [93] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," in *International Conference on Learning Representations (ICLR)*, 2020.
- [94] A. Zerroug, M. Vaishnav, J. Colin, S. Musslick, and T. Serre, "A benchmark for compositional visual reasoning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 29 776–29 788, 2022.
- [95] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, "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.
- [96] C. Zhang, B. Jia, S.-C. Zhu, and Y. Zhu, "Abstract spatial-temporal reasoning via probabilistic abduction and execution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9736–9746.
- [97] H. Zhang and T. Yu, "Alphazero," *Deep Reinforcement Learning: Fundamentals, Research and Applications*, pp. 391–415, 2020.
- [98] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," *arXiv preprint arXiv:2402.19473*, 2024.
- [99] C. Zhou, F. G. Redondo, J. Büchel, I. Boybat, X. T. Comas, S. Nandakumar, S. Das, A. Sebastian, M. Le Gallo, and P. N. Whatmough, "Ml-hw co-design of noise-robust tinymt models and always-on analog compute-in-memory edge accelerator," *IEEE Micro*, vol. 42, no. 6, pp. 76–87, 2022.
- [100] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, vol. 130, no. 9, pp. 2337–2348, 2022.