



# CoCoSys

CENTER FOR THE  
CO-DESIGN OF COGNITIVE SYSTEMS

---

## Poster Catalog

---

May 16-17, 2023



Semiconductor  
Research  
Corporation



# Table of Contents

CoCoSys Annual Review | May 16-17, 2023

<a href="#">Table of Contents</a>	2
<a href="#">CoCoSys PI Directory</a>	5
<a href="#">CoCoSys Themes &amp; Tasks</a>	7
<a href="#">POSTER SESSION 1</a>	9
<a href="#">Theme 1 Overview</a>	9
<a href="#">Theme 1 Presentation Details</a>	11
<a href="#">1.1 Amrit Nagarajan</a>	11
<a href="#">1.2 Anthony Thomas</a>	12
<a href="#">1.3 Matin Ghavamizadeh</a>	13
<a href="#">1.4 Matin Ghavamizadeh &amp; McCoy Becker</a>	13
<a href="#">1.5 Timur Ibrayev</a>	14
<a href="#">1.6 Vignesh Sundaresha</a>	15
<a href="#">1.7 Xiaofan Yu</a>	16
<a href="#">1.8 Zishen Wan &amp; Hanchen Yang</a>	16
<a href="#">1.9 Anirudh Sundar</a>	17
<a href="#">1.10 Benjamin Reichman</a>	18
<a href="#">1.11 Chaojian Li</a>	18
<a href="#">1.12 Christopher Kymm</a>	19
<a href="#">1.13 Samuel Spetalnick</a>	20
<a href="#">POSTER SESSION 2</a>	21
<a href="#">Theme 2 Overview</a>	21
<a href="#">Theme 2 Presentation Details</a>	23
<a href="#">2.1 Abhimanyu Bambhaniya</a>	23



<a href="#">2.2 Ahmed Hasssan</a>	<a href="#">24</a>
<a href="#">2.3 Gopikrishnan Raveendran Nair</a>	<a href="#">24</a>
<a href="#">2.4 Haoran You</a>	<a href="#">25</a>
<a href="#">2.5 Jeffrey Yu</a>	<a href="#">26</a>
<a href="#">2.6 McCoy Becker</a>	<a href="#">27</a>
<a href="#">2.7 Nealson Li</a>	<a href="#">28</a>
<a href="#">2.8 Raveesh Garg</a>	<a href="#">28</a>
<a href="#">2.9 Soonha Hwang</a>	<a href="#">29</a>
<a href="#">2.10 Sourjya Roy</a>	<a href="#">30</a>
<a href="#">2.11 Tianqi Zhang</a>	<a href="#">30</a>
<a href="#">2.12 Yang Zhao</a>	<a href="#">31</a>
<a href="#">2.13 Yonggan Fu</a>	<a href="#">32</a>
<a href="#">2.14 Kaining Zhou</a>	<a href="#">33</a>
<a href="#">2.15 Ashwin Bhat</a>	<a href="#">33</a>
<a href="#">2.16 Abhishek Moita</a>	<a href="#">34</a>
<a href="#">2.17 Tian Jin</a>	<a href="#">35</a>
<a href="#">2.18 Eric Atkinson</a>	<a href="#">36</a>
<a href="#">POSTER SESSION 3</a>	<a href="#">37</a>
<a href="#">Themes 3 and 4 Overview</a>	<a href="#">37</a>
<a href="#">Theme 3 Presentation Details</a>	<a href="#">39</a>
<a href="#">3.1 Akul Malhotra</a>	<a href="#">39</a>
<a href="#">3.2 Da Eun Shim</a>	<a href="#">40</a>
<a href="#">3.3 Imtiaz Ahmed</a>	<a href="#">40</a>
<a href="#">3.4 Kartik Prabhu</a>	<a href="#">41</a>
<a href="#">3.5 Piyush Kumar</a>	<a href="#">42</a>
<a href="#">3.6 Siri Narla</a>	<a href="#">42</a>
<a href="#">3.7 Wangxin He</a>	<a href="#">43</a>
<a href="#">3.8 Zhenyu Wang</a>	<a href="#">44</a>



<a href="#">Theme 4 Presentation Details</a>	45
<a href="#">4.9 Arghadip Das</a>	45
<a href="#">4.10 Chris Richardson</a>	46
<a href="#">4.11 Ruokai Yin</a>	47
<a href="#">4.12 Soumendu Ghosh</a>	48
<a href="#">4.13 Tyler Lizzo</a>	49
<a href="#">4.14 Sakshi Choudhary</a>	49
<a href="#">4.15 Sai Aparna Aketi</a>	50
<a href="#">4.16 Yuhang Li</a>	51
<a href="#">4.17 Zishen Wan</a>	52
<a href="#">4.18 Guangyu Jiang</a>	53



# CoCoSys PI Directory

NAME & AFFILIATION	ASSOCIATED TASKS	EMAIL
<b>Anand Raghunathan</b> Purdue University	3131.005, 3131.006, 3131.007, 3131.008	raghunathan@purdue.edu
<b>Anca Dragan</b> University of California, Berkeley	3131.013	anca@berkeley.edu
<b>Arijit Raychowdhury</b> Georgia Institute of Technology	3131.011, 3131.012, 3131.015, 3131.003, 3131.007, 3131.009	arijit.raychowdhury@ ece.gatech.edu
<b>Azad Naeemi</b> Georgia Institute of Technology	3131.01	an42@gatech.edu
<b>Bruno Olshausen</b> University of California, Berkeley	3131.002, 3131.003, 3131.004	baolshausen@berkeley.edu
<b>Jae-sun Seo</b> Arizona State University	3131.009, 3131.010, 3131.011, 3131.012	jaesun.seo@asu.edu
<b>James DiCarlo</b> Massachusetts Institute of Technology	3131.001	dicarlo@mit.edu
<b>Jan Rabaey</b> University of California, Berkeley	3131.002, 3131.007, 3131.008, 3131.005, 3131.006	jan@eecs.berkeley.edu
<b>Josh Tenenbaum</b> Massachusetts Institute of Technology	3131.001, 3131.004, 3131.013	jbt@mit.edu
<b>Kaushik Roy</b> Purdue University	3131.001, 3131.003, 3131.009, 3131.010	kaushik@purdue.edu
<b>Larry Heck</b> Georgia Institute of Technology	3131.013, 3131.014	larryheck@gatech.edu
<b>Michael Carbin</b> Massachusetts Institute of Technology	3131.006	mcarbin@csail.mit.edu



NAME & AFFILIATION	ASSOCIATED TASKS	EMAIL
<b>Naresh Shanbhag</b> University of Illinois at Urbana-Champaign	3131.001, 3131.002, 3131.003, 3131.004, 3131.015	shanbhag@illinois.edu
<b>Priyadarshini Panda</b> Yale University	3131.013, 3131.014, 3131.015	priya.panda@yale.edu
<b>Priyanka Raina</b> Stanford University	3131.006, 3131.009, 3131.012,	praina@stanford.edu
<b>Sumeet K. Gupta</b> Purdue University	3131.01	guptask@purdue.edu
<b>Tajana S. Rosing</b> University of California, San Diego	3131.002, 3131.005, 3131.015	tajana@ucsd.edu
<b>Tushar Krishna</b> Georgia Institute of Technology	3131.005, 3131.006, 3131.007, 3131.008	tushar@ece.gatech.edu
<b>Vijay Raghunathan</b> Purdue University	3131.014	vr@purdue.edu
<b>Yingyan (Celine) Lin</b> Georgia Institute of Technology	3131.006, 3131.009, 3131.012	yingyan.lin@rice.edu
<b>Yu (Kevin) Cao</b> Arizona State University	3131.007, 3131.008, 3131.011	ycao@asu.edu

## CoCoSys Support Staff

NAME	ROLE	EMAIL
<b>Emily Watson</b>	Program & Operations Manager	emily.watson@ece.gatech.edu
<b>Janna Young</b>	Faculty Support Coordinator	jyoung381@gatech.edu
<b>Melissa Donahoe</b>	Financial Analyst	melissa.donahoe@ ece.gatech.edu

# CoCoSys Themes & Tasks

**COCOSYS** aims to enable the next generation of collaborative human-AI systems through synergistic advances in algorithms, hardware motifs, algorithm-hardware co-design, and collective and collaborative intelligence. To pursue its overarching vision and goals, the center will adopt a vertically integrated approach consisting of synergistic efforts in neural, symbolic, and probabilistic algorithms; algorithm-hardware co-design; technology-driven hardware motifs; and collective and collaborative intelligence.

## THEME 1: NEURAL, SYMBOLIC, AND PROBABILISTIC ALGORITHMS

Theme 1 will create the next generation of explainable algorithms, expand the scope of neuro-inspired algorithms from perception to reasoning and decision-making, and uncover the fundamental accuracy-robustness-efficiency tradeoffs in cognitive systems.

Task Number	Task Name
3131.001	Unifying Neural, Symbolic and Probabilistic Models
3131.002	Hyper-dimensional (HD) Information Representations & Processing
3131.003	Computing with Emergent and Dynamical Systems
3131.004	Theoretical Underpinnings of Robustness-accuracy-efficiency Tradeoffs

## THEME 2: HARDWARE ALGORITHM CO-DESIGN

Theme 2 will distill the key computational characteristics of future cognitive workloads developed by Theme 1 and use them to drive the design of the next generation of programmable hardware architectures for cognitive computing. This theme will play a key role in ensuring that the developed algorithms are well-matched to the proposed hardware fabrics and vice-versa.

Task Number	Task Name
3131.005	Architectures for Neuro-symbolic-probabilistic Workloads
3131.006	Full-stack Optimization and Software Frameworks for Cognitive Systems
3131.007	Technology and Integration-driven Cognitive Architectures
3131.008	System Evaluation and Benchmarking



### THEME 3: TECHNOLOGY-DRIVEN HARDWARE MOTIFS

Theme 3 will design the building blocks of future cognitive hardware platforms by matching the unique capabilities of various CMOS and beyond-CMOS devices and integration technologies to the needs of the workloads, seeking quantum improvements in energy efficiency and performance.

Task Number	Task Name
3131.009	Digital, Mixed-signal and Mixed-mode Cognitive Circuits
3131.010	Technology (Logic, Memory, Interconnect) Evaluation
3131.011	Heterogeneous Integration Driven Cognitive HW Design
3131.012	Hardware Prototyping and Benchmarking

### THEME 4: COLLABORATIVE INTELLIGENCE

Theme 4 will specifically focus on the challenges involved in collections of AI agents and how AI agents interact with humans.

Task Number	Task Name
3131.013	Human-AI Collaboration Through Visual and Natural Language Understanding
3131.014	AI-AI Collaboration and Multi-agent Systems
3131.015	Robust, Secure and Privacy-preserving Intelligence





## POSTER SESSION I

# Theme I Overview

Tuesday, May 16 @ 11:10 AM-12:10 PM

### Zone I

POSTER NO.	PRESENTER	TITLE
I.1	Amrit Nagarajan	FASTRAIN-GNN: Fast and Accurate Self-Training for Graph Neural Networks
I.2	Anthony Thomas	A Formal Perspective on Learning with Hyperdimensional Computing
I.3	Matin Ghavamizadeh	Real-time Inference for Probabilistic Programs via Compute-Adaptive Sequential Monte Carlo
I.4	McCoy Becker & Matin Ghavamizadeh	Automatically Optimizing Probabilistic Programs for In-Distribution and Out-Of-Distribution Workloads
I.5	Timur Ibrayev	Two-Stream Active Learning with Foveation for Weakly-Supervised Object Localization
I.6	Vignesh Sundaresha	Comprehending the Adversarial Vulnerability of Randomized Ensembles
I.7	Xiaofan Yu	Lifelong, Federated and Multimodal Learning beyond the Edge with Hyperdimensional Computing
I.8	Zishen Wan & Hanchen Yang	Towards Cognitive AI System: A Survey and Prospective on Neuro-Symbolic AI



1.9	Anirudh Sundar	cTBLS: Augmenting Large Language Models with Conversational Tables
1.10	Benjamin Reichman	Directions for the Open-Knowledge Visual Question Answering Challenge
1.11	Chaojian Li	Instant-3D: Instant Neural Radiance Fields Training Towards Real-Time AR/VR 3D Reconstruction
1.12	Christopher Kymm	Residue Vector Symbolic Algebras Enable Efficient Neural Algorithms
1.13	Samuel Spetalnick	Probing Compute-in-Memory Design Constraints through 3 Generations of Taped-Out RRAM Macros

# Theme I Presentation Details

Tuesday, May 16 @ 11:10 AM-12:10 PM

Zone I

RESEARCH SCHOLAR	POSTER DETAILS
<p><b>I.I Amrit Nagarajan</b> Purdue University</p>  <p><b>Email:</b> <a href="mailto:nagaraj9@purdue.edu">nagaraj9@purdue.edu</a> <b>PI:</b> Anand Raghunathan <b>Expected Graduation Date:</b> December 2023</p>	<p><b>Title:</b> FASTRAIN-GNN: Fast and Accurate Self-Training for Graph Neural Networks</p> <p><b>Abstract:</b> Few-shot learning with GNNs is an important challenge in expanding their remarkable success. However, supervised training methods fail in few-shot learning. Self-training, wherein the GNN is trained in stages by augmenting the training data with a subset of the unlabeled data has emerged as a promising approach. However, self-training significantly increases the computational demands of GNN training. To address this challenge, we propose FASTRAIN-GNN, a framework for efficient and accurate few-shot learning. FASTRAIN-GNN performs four main optimizations: (1) Sampling-based Pseudolabel Filtering removes nodes whose pseudo labels are likely to be incorrect from the enlarged training set. (2,3) Dynamic Sizing and Dynamic Regularization find the optimal network architecture and amount of training regularization, and (4) Progressive Graph Pruning removes selected edges between nodes in the training set. FASTRAIN-GNN produces models that are consistently more accurate, while also substantially reducing the self-training time over the current state-of-the-art few-shot learning methods.</p> <p><b>Additional author(s):</b> Anand Raghunathan</p>



## I.2 Anthony Thomas

UC San Diego



**Email:** [ahthomas@eng.ucsd.edu](mailto:ahthomas@eng.ucsd.edu)

**PI:** Tajana Rosing

**Expected Graduation Date:**  
June 2023

**Title:** A Formal Perspective on Learning with Hyperdimensional Computing

**Abstract:** Hyperdimensional computing (HDC) is a broad family of techniques for expressing cognitive information processing tasks on high-dimensional and distributed representations of data. The choice of operators for embedding raw data and manipulating the high-dimensional representations is called a vector-symbolic architecture (VSA). VSAs provide a principled mechanism for integrating diverse types of input data, for instance, generated by different types of sensory systems, and representing it in a common format suitable for use in computationally efficient and neurally plausible learning algorithms. In this work, we study the performance of learning algorithms that run on VSA representations of data through the lens of kernel methods and statistical learning theory. We characterize the optimal solution to a wide range of learning tasks and analyze the tradeoff between the choice of encoding dimension, precision, and model performance. Our work also elucidates the connections between HDC and kernel methods, an influential and closely related area of research in the statistics and machine learning community.

**Additional author(s):** Sanjoy Dasgupta, Tara Javidi, Tajana Rosing



### I.3 Matin Ghavamizadeh

MIT



**Email:** [mghavami@mit.edu](mailto:mghavami@mit.edu)

**PI:** Josh Tenenbaum and  
Vikash Mansinghka

**Expected Graduation Date:**

June 2028

**Title:** Real-time Inference for Probabilistic Programs via Compute-Adaptive Sequential Monte Carlo

**Abstract:** We describe a family of Monte Carlo algorithms that can spend more time on surprising observations to produce high-quality inferences. Our approach relies on recent techniques that employ meta-inference to produce unbiased estimates of marginal likelihoods. We characterize and prove asymptotic soundness results for our family of algorithms and showcase our approach by a GPU implementation of a Bayesian 3D pose estimator. Preliminary results suggest this approach may enable our pose estimator to compete in speed with neural networks while competing in accuracy with state-of-the-art Bayesian inverse graphics.

**Additional author(s):** McCoy Becker, Nishad Gothoskar, Cameron Freer, Vikash Mansinghka

### I.4 Matin Ghavamizadeh & McCoy Becker

MIT



**Email:** [mghavami@mit.edu](mailto:mghavami@mit.edu) and  
[mccoyb@mit.edu](mailto:mccoyb@mit.edu)

**PI:** Josh Tenenbaum and Vikash  
Mansinghka


**Expected Graduation Date:**

June 2028 and June 2027,  
respectively

**Title:** Automatically Optimizing Probabilistic Programs for In-Distribution and Out-Of-Distribution Workloads

**Abstract:** We present a method for jointly tuning the parameters of a generative model and an inference algorithm to optimize performance on data generated from a “workload” probabilistic program. Crucially, the workload distribution need not be the same distribution assumed by the generative model. We achieve this by stochastic optimization of the symmetric conditional relative entropy between the sampling distribution of the inference algorithm and workload posterior. We provide an efficient GPU implementation that works with the GenJAX probabilistic programming system. We also show how to reduce estimator variance and tuning cost for domain-specific PPLs embedded in GenJAX, by synthesizing specialized estimation algorithms that exploit the structure in those embedded domain-specific PPLs. We illustrate the performance gains from tuning using examples such as hidden Markov modeling and real-time 3D object tracking.



	<p><b>Additional author(s):</b> Matin Ghavami, Nishad Gothoskar, Martin Rinard, Vikash Mansinghka</p>
<p><b>I.5 Timur Ibrayev</b> Purdue University</p>  <p><b>Email:</b> tibrayev@purdue.edu <b>PI:</b> Kaushik Roy <b>Expected Graduation Date:</b> June 2024</p>	<p><b>Title:</b> Two-Stream Active Learning with Foveation for Weakly-Supervised Object Localization</p> <p><b>Abstract:</b> Current machine perception frameworks process the entire input in a one-shot manner to provide answers to both “what object is being observed” and “where it is located”. Contrary, the “two-stream hypothesis” describes the neural processing in the visual cortex as an active vision system that utilizes two separate regions of the brain. Hence, we propose a machine learning framework that models the mechanisms found in ventral (what) and dorsal (where) streams and explore the potential benefits that it offers. By training the proposed framework using label-based supervised training for the ventral stream model and using reinforcement learning for the dorsal stream model, we show that it is applicable to the challenging task of weakly-supervised object localization (WSOL). Specifically, the framework is capable of predicting both the properties of an object and its bounding box location, while being limited to only object class or its attributes during training.</p> <p><b>Additional author(s):</b> Amitangshu Mukherjee, Sai Aparna Aketi, Kaushik Roy</p>



## I.6 Vignesh Sundaresha UIUC



**Email:** vs49@illinois.edu

**PI:** Naresh Shanbhag

**Expected Graduation Date:**  
December 2027

**Title:** Comprehending the Adversarial Vulnerability of Randomized Ensembles

**Abstract:** Despite the tremendous success of deep neural networks across various tasks, their vulnerability to imperceptible adversarial perturbations has hindered their deployment in the real world. Recently, works on randomized ensembles have empirically demonstrated significant improvements in adversarial robustness over standard adversarially trained (AT) models with minimal computational overhead, making them a promising solution for safety-critical resource-constrained applications. However, this impressive performance raises the question: Are these robustness gains provided by randomized ensembles real? In this work we address this question both theoretically and empirically. We first establish theoretically that commonly employed robustness evaluation methods such as adaptive PGD provide a false sense of security in this setting. Subsequently, we propose a theoretically-sound and efficient adversarial attack algorithm (ARC) capable of compromising random ensembles even in cases where adaptive PGD fails to do so. We conduct comprehensive experiments across a variety of network architectures, training schemes, datasets, and norms to support our claims, and empirically establish that randomized ensembles are in fact more vulnerable to  $\ell_p$ -bounded adversarial perturbations than even standard AT models

**Additional author(s):** Hassan Dbouk, Naresh Shanbhag



## I.7 Xiaofan Yu

UC San Diego



**Email:** [xlyu@eng.ucsd.edu](mailto:xlyu@eng.ucsd.edu)

**PI:** Tajana Rosing

**Expected Graduation Date:**  
June 2024

**Title:** Lifelong, Federated and Multimodal Learning beyond the Edge with Hyperdimensional Computing

**Abstract:** With the recent advancements in machine learning algorithms and powerful edge computing platforms, both training and inference at the edge are possible. On-device training dramatically reduces the total amount of data that has to be sent via the network, thus reducing the overall energy costs, but it also provides for much faster decision-making at the edge. However, multiple challenges exist in real edge deployment on (1) how to effectively learn from drifting, distributed, and multimodal data from sensors, and (2) how to perform efficient computation subject to limited resources and energy. Novel lightweight algorithms are needed. We introduce our research on enabling lifelong, federated, and multimodal learning with Hyperdimensional Computing for low-power edge devices. Our method delivers comparable (if not better) accuracy performance compared to state-of-the-art ML models such as LSTM while being 3x-10x more efficient in terms of time and energy.

**Additional author(s):** Quanling Zhao, Anthony Thomas, and Tajana Rosing

## I.8 Zishen Wan & Hanchen Yang

Georgia Tech



**Email:** [zishenwan@gatech.edu](mailto:zishenwan@gatech.edu)  
and [hanchen@gatech.edu](mailto:hanchen@gatech.edu)

**PI:** Arijit Raychowdhury &  
Tushar Krishna

**Title:** Towards Cognitive AI System: A Survey and Prospective on Neuro-Symbolic AI

**Abstract:** The remarkable advancements in artificial intelligence (AI), primarily driven by deep neural networks, have significantly impacted various aspects of our lives. However, the current challenges surrounding unsustainable computational trajectories, limited robustness, and a lack of explainability call for the development of next-generation AI systems. Neuro-symbolic AI (NSAI) emerges as a promising paradigm, fusing neural, symbolic, and probabilistic approaches to enhance interpretability, robustness, and trustworthiness while facilitating learning from much less data. Recent NSAI systems have demonstrated great potential in collaborative human-AI scenarios with reasoning and cognitive capabilities. In this presentation, we provide a systematic review of





<p><b>Expected Graduation Date:</b> May 2025 and December 2027, respectively</p>	<p>recent progress in NSAI and analyze the performance characteristics and computational operators of NSAI models. Furthermore, we discuss the challenges and potential future directions of NSAI from both system and architectural perspectives.</p> <p><b>Additional author(s):</b> Zishen Wan, Che-Kai Liu, Hanchen Yang, Chaojian Li, Haoran You, Yonggan Fu, Cheng Wan, Tushar Krishna, Yingyan (Celine) Lin, Arijit Raychowdhury</p>
<p><b>I.9 Anirudh Sundar</b> Georgia Tech</p>  <p><b>Email:</b> asundar34@gatech.edu <b>PI:</b> Larry Heck <b>Expected Graduation Date:</b> December 2025</p>	<p><b>Title:</b> cTBLS: Augmenting Large Language Models with Conversational Tables</p> <p><b>Abstract:</b> An open challenge in multimodal conversational AI requires augmenting language models with information from textual and non-textual sources for multi-turn dialogue. In contrast to prior work addressing Table Question Answering and Knowledge Grounded Response Generation as separate tasks, we develop Conversational Tables (cTBLS), a three-step encoder-decoder architecture to retrieve tabular information and generate dialogue responses end-to-end. cTBLS uses Transformer encoder embeddings for Dense Table Retrieval and obtains up to 5% relative improvement in Top-1 and Top-3 accuracy over sparse retrieval on the HybriDialogue dataset. Additionally, cTBLS performs tabular knowledge retrieval using encoder and decoder models, resulting in up to 46% relative improvement in ROUGE scores and better human evaluation for response generation on HybriDialogue. While cTBLS currently prompts Very Large Language Models (175B parameters) for response generation, future work will assess the feasibility of fine-tuning order-of-magnitude smaller Large Language Models (13B parameters) for deployment in resource-constrained environments.</p> <p><b>Additional author(s):</b> Larry Heck</p>



**I.10 Benjamin Reichman**  
Georgia Tech



**Email:** [bzr@gatech.edu](mailto:bzr@gatech.edu)

**PI:** Larry Heck

**Expected Graduation Date:**  
June 2026

**Title:** Directions for the Open-Knowledge Visual Question Answering Challenge

**Abstract:** Visual question answering (VQA) lies at the intersection of language and vision research. It functions as a building block for multimodal conversational AI and serves as a testbed for assessing a model's capability for open-domain scene understanding. The 2019 Outside Knowledge VQA dataset "OK-VQA" extends the VQA task by adding more challenging questions that require complex, factual, and commonsense knowledge. However, in our analysis, we found that 41.4% of the dataset needed to be corrected and 10.6% needed to be removed. Clean datasets make the training of AI models less noisy and provide better testbeds for analyzing models. Using the improved dataset, we analyzed the state-of-the-art systems that address this task. We found common problems in how knowledge retrieval is performed and how this knowledge is then translated into a response. Our current work focuses on improving these AI systems by improving performance on both subproblems.

**Additional author(s):** Larry Heck

**I.11 Chaojian Li**  
Georgia Tech



**Email:** [cli851@gatech.edu](mailto:cli851@gatech.edu)

**PI:** Yingyan (Celine) Lin

**Title:** Instant-3D: Instant Neural Radiance Fields Training Towards Real-Time AR/VR 3D Reconstruction

**Abstract:** Neural Radiance Field (NeRF) based 3D reconstruction has promised to enable numerous emerging immersive applications in Augmented and Virtual Reality (AR/VR). However, while it is highly desired to unleash the above promise, instant (i.e., less than 5 seconds) on-device NeRF-based 3D reconstruction is still not possible on AR/VR devices even when using the most efficient NeRF training algorithm. In this work, we propose an algorithm-hardware co-design acceleration framework called Instant-3D, which to the best of our knowledge is the first to achieve instant on-device NeRF training for AR/VR. Excitingly, our Instant-3D has fulfilled the goal of instant 3D reconstruction for AR/VR, requiring only 1.6 seconds per scene while meeting the AR/VR power consumption constraint of 1.9 W.



<p><b>Expected Graduation Date:</b> December 2024</p>	<p><b>Additional author(s):</b> Sixu Li, Wenbo Zhu, Boyang (Tony) Yu, Yang (Katie) Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan (Celine) Lin</p>
<p><b>I.12 Christopher Kymm</b> UC Berkeley</p>  <p><b>Email:</b> <a href="mailto:cjkymn@berkeley.edu">cjkymn@berkeley.edu</a> <b>PI:</b> Bruno Olshausen <b>Expected Graduation Date:</b> June 2025</p>	<p><b>Title:</b> Residue Vector Symbolic Algebras Enable Efficient Neural Algorithms</p> <p><b>Abstract:</b> A fundamental problem in neural computation is the effective representation and transformation of variables by use of high-dimensional vector spaces (e.g., population codes). The desiderata include high storage capacity, efficient computation over, and robustness to noise. A promising population of neurons that fulfill these criteria are grid cells in medial entorhinal cortex, which theory suggests could implement a residue number system (RNS) for spatial position. Here, we describe how the properties of residue number systems implemented via distributed representations can be effectively used in algorithms utilizing random, high-dimensional vectors (Vector Symbolic Architectures/VSA). VSA implementations of RNS generalize the power of the encoding from integers to real values, allow robust decoding, and enable algorithms for combinatorial search problems. Conversely, RNS improves the expected space and time complexity of VSA algorithms. We provide empirical and analytic support quantifying the capacity of our encoding scheme and report performance on multiple applications.</p> <p><b>Additional author(s):</b> Denis Kleyko, Connor Bybee, E. Paxon Frady, Pentti Kanerva, Friedrich T. Sommer, Bruno A. Olshausen</p>



## I.13 Samuel Spetalnick

Georgia Tech



**Email:** sspetalnick3@gatech.edu

**PI:** Arijit Raychowdhury

**Expected Graduation Date:**

December 2023

**Title:** Probing Compute-in-Memory Design Constraints through 3 Generations of Taped-Out RRAM Macros

**Abstract:** Compute-in-memory (CIM) captures the concept of performing multiply accumulate (MAC) operations in a nonvolatile resistive memory array by summing the current contributed by low-resistance-state (LRS) cells representing stored matrix elements. This conceptually clean primitive might offer advantages in energy efficiency or normalized bandwidth relative to traditional (digital) systems. Despite recent work in this area, practical concerns about accuracy, memory density, and net efficiency for dense vectors have persisted. We present the evolution of a current-sensing 40nm CIM macro using a foundry process across three design generations, discussing the key innovations and challenges at each step toward improving the accuracy, density, and scalability of the macro. The latest work introduces hybrid data-aware analog and mixed-signal offset cancellation to counter channel mismatch and off-state cell current, pseudo-4-terminal BL/SL regulation to address  $I_{\text{CELL}} R_{\text{BLSL}}$  and  $R_{\text{MUX}}$  drop, and a low  $V_{\text{CELL}}$  bias point to improve functionality with dense weight matrices.

**Additional author(s):** Muya Chang, Brian Crafton, Ashwin S. Lele, Shota Konno, Arijit Raychowdhury



## POSTER SESSION 2

# Theme 2 Overview

Tuesday, May 16 @ 2:15-3:15 PM

**Zone 2**

POSTER NO.	PRESENTER	TITLE
2.1	Abhimanyu Bambhaniya	Understanding Trade-Offs between Area, Performance, and Quality for Architecting Sparse Attention Accelerators
2.2	Ahmed Hasssan	QSNN: A Memory and Energy Efficient Low Precision Spiking Neural Network for the Computer Vision
2.3	Gopikrishnan Raveendran Nair	Reconfigurable Accelerator Design for Heterogeneous AI Models
2.4	Haoran You	EyeCoD: Eye Tracking System Acceleration via FlatCam-based Algorithm & Accelerator Co-Design
2.5	Jeffrey Yu	Transformer Inference and Fine-tuning on Resource-Constrained Edge Devices
2.6	McCoy Becker	Scaling Probabilistic Programming via GenJAX and OpenGen
2.7	Nealson Li	Eye Tracking with Event Camera for Extended Reality (XR) Applications
2.8	Raveesh Garg	Flexagon: A Multi-Dataflow Sparse-Sparse Matrix Multiplication Accelerator for Efficient DNN Processing



2.9	Soonha Hwang	Proper Benchmarking In-Memory Computing Architectures
2.10	Sourjya Roy	Evaluation of Spintronic Memory in Machine Learning Training Accelerators
2.11	Tianqi Zhang	HyperSpikeASIC: Accelerating Event-based Workloads with HyperDimensional Computing and Spiking Neural Networks
2.12	Yang Zhao	Instant-NeRF: Instant On-Device Neural Radiance Field Training via Algorithm-Accelerator Co-Designed Near-Memory Processing
2.13	Yonggan Fu	Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design
2.14	Kaining Zhou	IMCsim: In-Memory Computing Simulator
2.15	Ashwin Bhat	Hardware-Software Co-Design to Enable Low-Latency Explainable AI
2.16	Abhishek Moita	XPer: Peripheral Circuit & Neural Architecture Co-search for Area and Energy-efficient Xbar-based Computing
2.17	Tian Jin	Pruning's Effect on Generalization Through the Lens of Training and Regularization
2.18	Eric Atkinson	A Language and Logic for Programming and Reasoning with Partial Observability

# Theme 2 Presentation Details

Tuesday, May 16 @ 2:15-3:15 PM

Zone 2

RESEARCH SCHOLAR	POSTER DETAILS
<p><b>2.1 Abhimanyu Bambhaniya</b> Georgia Tech</p>  <p><b>Email:</b> abambhaniya3@gatech.edu <b>PI:</b> Tushar Krishna <b>Expected Graduation Date:</b> May 2025</p>	<p><b>Title:</b> Understanding Trade-Offs between Area, Performance, and Quality for Architecting Sparse Attention Accelerators</p> <p><b>Abstract:</b> The constantly increasing size of sparse neural networks presents a unique opportunity for designing efficient accelerators that can offer speedup on sparse workloads. Enhancing an accelerator to support sparsity requires improvement to its three primary building blocks: compute, memory, and interconnect. However, there is a lack of systematic methodology to provide architectural insights to design efficient sparse accelerators for variants of sparse neural networks. In this work, we systematically analyze the trade-offs between area and performance contingent on a target quality constraint across a diverse range of attention networks. We propose a methodology to quantify the area overhead of major accelerator components (e.g. compute, memory, and interconnect) to enable sparse execution of attention networks, and we develop an analytical cost model, MODESA, that estimates the area and performance for attention-based deep neural networks and hardware. Finally, we demonstrate how MODESA enables identifying efficient sparse accelerator configurations for attention-based models.</p> <p><b>Additional author(s):</b> Sheng-Chun Kao, Geonhwa Jeong, Amir Yazdanbakhsh, Suvinay Subramanian, Tushar Krishna</p>



## 2.2 Ahmed Hassan ASU



**Email:** ahassan@asu.edu

**PI:** Jae-sun Seo

**Expected Graduation**

**Date:** May 2025

**Title:** QSNN: A Memory and Energy Efficient Low Precision Spiking Neural Network for the Computer Vision

**Abstract:** Spiking Neural Networks (SNNs) are emerging as an inexpensive alternative to Deep Neural Networks (DNNs) due to their low energy consumption, latency, and computational complexity in forward propagation. Binary spike-based information transmission reduces memory and power consumption as compared to analog values in DNNs. However, the convolution operations in the intermediate layers of the network generate high precision membrane potential that can increase memory occupancy, energy consumption, and latency on the software and hardware level. To further reduce the SNNs computational cost, we propose QSNN reducing the high precision membrane potential and weights to the fixed 2-bit and 4-bit precision. We use a sigmoid approximation in the backward pass to overcome the gradient mismatch at the quantization stage. We benchmark our algorithm on multiple SNN-based architectures using Dynamic Vision Sensors (DVS) and static image datasets. The proposed work demonstrates high performance and surpasses the existing state-of-the-art works with 5x less storage occupancy and energy consumption.

**Additional author(s):** Jian Meng, Jae-sun Seo





## 2.3 Gopikrishnan Raveendran Nair

ASU



**Email:** graveenl@asu.edu

**PI:** Yu (Kevin) Cao

**Expected Graduation**

**Date:** December 2024

**Title:** Reconfigurable Accelerator Design for Heterogeneous AI Models

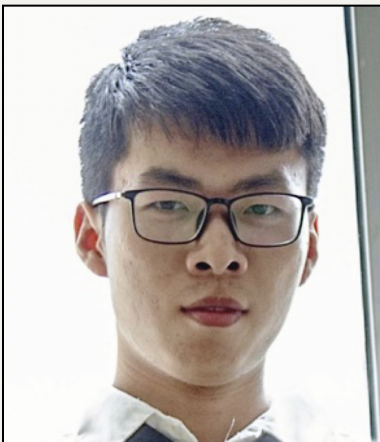
**Abstract:** The field of AI has expanded beyond DNNs to include GCNs and transformers, which differ in model size, dataflow, memory access patterns, and data sparsity. Unfortunately, most of the current hardware platforms like GPUs and ASICs are optimized for regular DNNs, making them inefficient for handling massive, unstructured, and sparse GCNs. Even in-memory computing (IMC), which has a parallel and homogeneous architecture suitable for dense matrix/vector computation, is underutilized for sparse and random data patterns.

In this work, we propose a reconfigurable accelerator that balances the computational needs of various AI models by analyzing DNNs and GCNs and identifying common design macros. Reconfigurability is introduced at computation and data movement levels, with latch-based digital IMCs, SIMD processing elements, and reconfigurable scatter/gather and buffer modules. A cycle-accurate simulator benchmarks performance, with successful mapping of DNN models (VGG16, ResNet) and GCN models on the datasets Cora and Citeseer, with the marginal area and latency overhead.

**Additional author(s):** Yu Cao

## 2.4 Haoran You

Georgia Tech




**Title:** EyeCoD: Eye Tracking System Acceleration via FlatCam-based Algorithm & Accelerator Co-Design

**Abstract:** Eye tracking has become an essential human-machine interaction modality in virtual and augmented reality (VR/AR) applications requiring high throughput (e.g., >240 FPS), small form-factor, and enhanced visual privacy. Existing eye tracking systems have adopted bulky lens-based cameras and thus suffer from both a large form-factor and high communication cost between the camera and backend processor. This work presents a camera, algorithm, and accelerator co-designed lensless eye-tracking system dubbed EyeCoD, which to the best of our knowledge is the first to provide a general



<p><b>Email:</b> <a href="mailto:haoran.you@gatech.edu">haoran.you@gatech.edu</a> <b>PI:</b> Yingyan (Celine) Lin <b>Expected Graduation Date:</b> December 2024</p>	<p>front-end eye-tracking solution for AR/VR while satisfying the requirements for both high throughput and smaller form-factor. Specifically, EyeCoD integrates system-, algorithm-, and accelerator-level techniques to boost system efficiency without sacrificing eye-tracking accuracy. We believe that our EyeCoD system will pave the way for next-generation eye-tracking solutions in VR/AR and can shed light on future innovations for intelligent imaging systems.</p> <p><b>Additional author(s):</b> Cheng Wan, Yang Zhao, Zhongzhi Yu, Yonggan Fu, Jiayi Yuan, Shang Wu, Shunyao Zhang, Yongan Zhang, Chaojian Li, Vivek Boominathan, Ashok Veeraraghavan, Ziyun Li, Yingyan Lin</p>
<p><b>2.5 Jeffrey Yu</b> Stanford University</p>  <p><b>Email:</b> <a href="mailto:jeffrey@stanford.edu">jeffrey@stanford.edu</a> <b>PI:</b> Priyanka Raina <b>Expected Graduation Date:</b> June 2027</p>	<p><b>Title:</b> Transformer Inference and Fine-tuning on Resource-Constrained Edge Devices</p> <p><b>Abstract:</b> Implementing Transformer inference and fine-tuning on edge devices presents a significant challenge due to limited computing resources and memory capacity. This task becomes even more difficult with the emergence of larger and deeper networks, such as BERT and GPT-3. To address this issue, we have developed a methodology that performs Transformer inference and fine-tuning using an 8-bit datatype called Posit (Posit8) on memory-constrained edge accelerators without compromising accuracy or modifying model architectures. Our approach involves storing activation, weights, and gradients in Posit8 while keeping bias and bias gradients in 16-bit Posit (Posit16), which reduces the memory footprint by almost 50% compared to the 16-bit floating-point. To prevent training accuracy degradation caused by the narrower range of Posit8, we propose three techniques. Firstly, we fuse matrix multiplication with element-wise operations such as residual addition and attention scaling to reduce quantization error during inference. Second, we apply automatic loss scaling to shift activation gradients into the Posit8 representable range. Finally, we propose stochastic gradient descent (SGD) with weight-splitting, a technique that allows weights</p>

	<p>and weight gradients to stay in Posit8 while training to full accuracy as the 16-bit floating point. Our methodology achieves MobileBERT inference and fine-tuning accuracy on GLUE tasks equivalent to that of 16-bit floating-point models, while using only 50% of the training memory, making it a highly efficient approach for implementing deep neural network training and inference on edge devices with limited computing resources and memory capacity.</p> <p><b>Additional author(s):</b> Kartik Prabhu, Yonaton Urman</p>
<p><b>2.6 McCoy Becker</b> MIT</p>  <p><b>Email:</b> mccoymb@mit.edu <b>PI:</b> Josh Tenenbaum and Vikash Mansingha <b>Expected Graduation Date:</b> June 2027</p>	<p><b>Title:</b> Scaling Probabilistic Programming via GenJAX and OpenGen</p> <p><b>Abstract:</b> We present a software ecosystem centered on the Gen probabilistic programming language designed to scale probabilistic programming to the next generation of hardware acceleration. We illustrate applications of Gen to automate common sense and perceptual reasoning tasks with state-of-the-art performance, as well as scientific modeling in neuroscience. We showcase accelerated versions of Gen optimized to target CPU and resource-limited devices (GenTorch / GenTL), as well as TPUs and other parallel fabrics (GenJAX). We showcase benchmark timings across the ecosystem of Gen implementations.</p> <p><b>Additional author(s):</b> Marco Cusumano-Towner, Matin Ghavami, Nishad Gothoskar, Cameron Freer, Vikash Mansingha</p>



## 2.7 Neelson Li

Georgia Tech



**Email:** [neelson@gatech.edu](mailto:neelson@gatech.edu)

**PI:** Arijit Raychowdhury

**Expected Graduation**

**Date:** May 2025

**Title:** Eye Tracking and Gaze Estimation with Event Camera

**Abstract:** Near-eye tracking and gaze estimation is a task that maps the recording of an eye captured by an adjacent camera to the direction of a person's gaze in space. Unlike frame-based cameras, event cameras generate asynchronous sparse events with high temporal resolution that are well-suited for recording fast eye movements. However, due to the natural differences in the data characteristics, designing algorithms and systems for event-based data is challenging. By analyzing eye parts and movements, and harnessing the polar, spatial, and temporal distribution of the events, we introduce two real-time pipelines to extract pupil features. Then, we present a recurrent neural network with a proposed coordinate-to-angle loss function to accurately estimate gaze from pupil feature sequence. Our system locates the pupil with an error of 3.68 pixels at 160 mW system power and estimates gaze with an angular accuracy of  $0.46^\circ$  and update rates of 950 Hz.

**Additional author(s):** Ashwin Bhat, Arijit Raychowdhury

## 2.8 Raveesh Garg

Georgia Tech



**Email:** [raveesh.g@gatech.edu](mailto:raveesh.g@gatech.edu)

**PI:** Tushar Krishna


**Expected Graduation**

**Date:** May 2025

**Title:** Flexagon: A Multi-Dataflow Sparse-Sparse Matrix Multiplication Accelerator for Efficient DNN Processing

**Abstract:** Sparsity is a growing trend in modern DNN models. Existing Sparse-Sparse Matrix Multiplication (SpMSPM) accelerators are tailored to a particular SpMSPM dataflow (i.e., Inner Product, Outer Product, or Gustavson's). We demonstrate that fixed dataflow results in a suboptimal solution since different SpMSPM kernels show varying features (i.e., dimensions, sparsity pattern, sparsity degree), which makes each dataflow better suited to different layers. In this work, we present Flexagon, the first SpMSPM reconfigurable accelerator that is capable of executing the dataflow that best matches each case. Flexagon accelerator is based on a novel Merger-Reduction Network (MRN) that unifies the concept of reducing and merging in the same substrate, increasing efficiency. We show that Flexagon achieves average performance benefits of 4.59x, 1.71x, and 1.35x with respect to the state-of-the-art



	<p>SIGMA-like(inner-product), Sparch-like(outer-product) and GAMMA-like(Gustavson's) accelerators (265%, 67% and 18%, respectively, in terms of average performance/area).</p> <p><b>Additional author(s):</b> Francisco Munoz-Martinez, Michael Pellauer, Jose L Abellan, Manuel E Acacio, Tushar Krishna</p>
<p><b>2.9 Soonha Hwang</b> UIUC</p>  <p><b>Email:</b> soonhah2@illinois.edu <b>PI:</b> Naresh Shanbhag <b>Expected Graduation Date:</b> May 2028</p>	<p><b>Title:</b> Proper Benchmarking In-Memory Computing Architectures</p> <p><b>Abstract:</b> In-memory computing (IMC) architectures have emerged as a compelling platform to implement energy-efficient machine learning (ML) systems. However, the energy efficiency gains provided by IMC designs seem to be leveling off and the limiting factors are not clear due to the conceptual complexity and a lack of rigorous benchmarking methodology for IMCs. To address this issue, we developed a benchmarking methodology supported by an extensive database of &gt;70 IMC designs published since 2018, to identify trends in this area. Our benchmarking effort (<a href="https://github.com/naresh-shanbhag/UIUC-IMC-Benchmarking">https://github.com/naresh-shanbhag/UIUC-IMC-Benchmarking</a>) indicates: 1) SRAM-based IMCs show a clear win in terms of energy efficiency and compute density over digital accelerators at the bank level, although this advantage diminishes when compared at the processor level; and 2) eNVM-based IMCs have lower energy efficiency and compute density than SRAM-based IMCs and, surprisingly lag digital accelerators in terms of compute density.</p> <p><b>Additional author(s):</b> Soonha Hwang, Saion K. Roy, and Naresh R. Shanbhag</p>



## 2.10 Sourjya Roy

Purdue University



**Email:** roy48@purdue.edu

**PI:** Anand Raghunathan

**Expected Graduation**

**Date:** May 2024

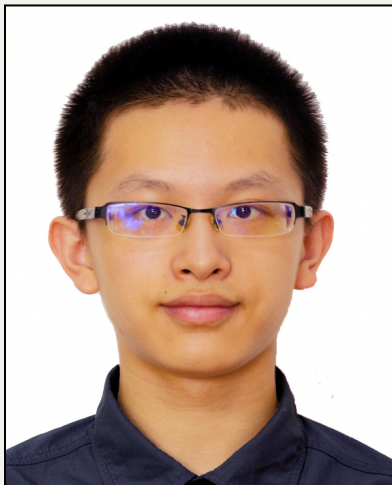
**Title:** Evaluation of Spintronic Memory in Machine Learning Training Accelerators

**Abstract:** Training leads to many off-chip memory accesses if the on-chip memory is not large enough to hold the required data structures. Among the emerging NVMs, STT-MRAM offers many desirable properties such as high endurance with reasonable access time. Compared to SRAM, STT-MRAM provides 3-4x higher density with substantially reduced leakage power. However, STT-MRAM requires higher write energy and time compared to CMOS-based SRAM, due to the use of a higher write voltage and longer write duration. This work performs a device-to-system evaluation and co-optimization of STT-MRAM for efficient ML training accelerator design. We propose to exploit MRAM with reduced write voltage and duration to address the large cost of high write operations and evaluate the impact of the low-cost write operations on both the system-level hardware performance and the accuracy of training DNN models. We achieve an order-of-magnitude improvement in energy efficiency relative to iso-capacity and iso-area SRAM baselines.

**Additional author(s):** Cheng Wang, Anand Raghunathan

## 2.11 Tianqi Zhang

UCSD




**Email:** tiz014@ucsd.edu

**PI:** Tajana Rosing

**Title:** HyperSpikeASIC: Accelerating Event-based Workloads with HyperDimensional Computing and Spiking Neural Networks

**Abstract:** Two emerging neuromorphic computing models are Hyperdimensional Computing (HDC) and Spiking Neural Networks (SNNs), both with their own benefits. HDC has various desirable properties that other ML algorithms lack such as robustness to noise, simple operations, and high parallelism. SNNs are able to process event-based signal data in an efficient manner. This work develops HyperSpike, which utilizes a single, randomly initialized, and untrained SNN layer as a feature extractor connected to a trained HDC classifier. HDC is used to enable more efficient classification and provide robustness to errors. We show that HyperSpike is 31.5x more robust to errors than SNNs. We further develop HyperSpikeASIC, a customized accelerator for HyperSpike. By



<p><b>Expected Graduation</b> <b>Date:</b> June 2026</p>	<p>decoupling the neuron and synapses, HyperSpikeASIC skips the inactive neurons and limits the neuron state updating to once per time step at most.</p> <p><b>Additional author(s):</b> Tianqi Zhang, Justing Morris, Kenneth Stewart, Hin Wai Lui, Behnam Khaleghi, Anthony Thomas, Thiago Goncalves-Marback, Baris Aksanli, Emre O. Neftci, Tajana Rosing</p>
<p><b>2.12 Yang Zhao</b> Rice University</p>  <p><b>Email:</b> zy34@rice.edu <b>PI:</b> Yingyan (Celine) Lin</p>	<p><b>Title:</b> Instant-NeRF: Instant On-Device Neural Radiance Field Training via Algorithm-Accelerator Co-Designed Near-Memory Processing</p> <p><b>Abstract:</b> Instant on-device Neural Radiance Fields (NeRFs) are in growing demand to unleash the promise of immersive AR/VR experiences, but still limited by their prohibitive training time. Our profiling analysis unveils a memory-bound bottleneck in NeRF training. To tackle this bottleneck, near-memory processing (NMP) promises to be an effective solution but also faces various challenges due to the unique workloads of NeRFs, including random hash table lookup, random point processing sequence, and heterogeneous bottleneck steps. Therefore, we propose the first NMP framework, Instant-NeRF, dedicated to enabling instant on-device NeRF training. Experiments on eight datasets consistently validate the effectiveness of Instant-NeRF.</p> <p><b>Additional author(s):</b> Shang Wu, Jingqun Zhang, Sixu Li, Chaojian Li, and Yingyan (Celine) Lin</p>



## 2.13 Yonggan Fu

Georgia Tech



**Email:** [yfu314@gatech.edu](mailto:yfu314@gatech.edu)

**PI:** Yingyan (Celine) Lin

**Expected Graduation**

**Date:** May 2025

**Title:** Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design

**Abstract:** Novel view synthesis is an essential functionality for enabling immersive experiences in various Augmented- and Virtual-Reality (AR/VR) applications, for which generalizable Neural Radiance Fields (NeRFs) have gained increasing popularity thanks to their cross-scene generalization capability. Despite their promise, the real-device deployment of generalizable NeRFs is bottlenecked by their prohibitive complexity due to the required massive memory accesses to acquire scene features, causing their ray marching process to be memory-bounded. To this end, we propose Gen-NeRF, an algorithm-hardware co-design framework dedicated to generalizable NeRF acceleration, which for the first time enables real-time generalizable NeRFs. On the algorithm side, Gen-NeRF integrates a coarse-then-focus sampling strategy, leveraging the fact that different regions of a 3D scene contribute differently to the rendered pixel, to enable sparse yet effective sampling. On the hardware side, Gen-NeRF highlights an accelerator micro-architecture to maximize the data reuse opportunities among different rays by making use of their epipolar geometric relationship. Furthermore, our Gen-NeRF accelerator features a customized dataflow to enhance data locality during point-to-hardware mapping and an optimized scene feature storage strategy to minimize memory bank conflicts.

**Additional author(s):** Zhifan Ye, Jiayi Yuan, Shunyao Zhang, Sixu Li, Haoran You, Yingyan (Celine) Lin





## 2.14 Kaining Zhou

UIUC



**Email:** kainingz@illinois.edu

**PI:** Naresh Shanbhag

**Expected Graduation**

**Date:** May 2027

**Title:** IMCsim: An In-Memory Computing Architecture Simulator

**Abstract:** In recent years, In-Memory Computing (IMC) has garnered increasing attention due to its high parallelism, energy efficiency, and computing density. However, these remarkable benefits have been demonstrated at a bank level. Specifically, IMC-based multi-bank processors have not yet demonstrated clear wins with respect to digital accelerators when complete AI workloads are mapped. This is because of the huge design space composed from the outer product of the IMC design space and the space of application-to-architecture mapping. To address this issue, a key first step is to develop a circuit-aware IMC architecture simulator. This work introduces IMCsim, an IMC simulation platform that encompasses cycle-accurate and full-system simulation, circuit-level behavior tracing, multibank and network-on-chip modeling, and memory access profiling. Furthermore, IMCsim is capable of supporting diverse IMC kernels, providing insights into its computing accuracy, and is programmable.

**Additional author(s):** Jian Huang, Nam-Sung Kim, Naresh Shanbhag

## 2.15 Ashwin Bhat

Georgia Tech



**Email:**

ashwinbhat@gatech.edu


**PI:** Arijit Raychowdhury

**Title:** Hardware-Software Co-Design to Enable Low-Latency Explainable AI

**Abstract:** There has been a surge in Explainable-AI (XAI) methods that provide insights into the workings of Deep Neural Network (DNN) models. However, compared to a single forward-pass inference, there is a significant computational overhead to generate the explanation which hinders real-time XAI. In this work, we target this issue for two popular post-hoc explanation methods namely (1) Integrated Gradients and (2) GradCAM.

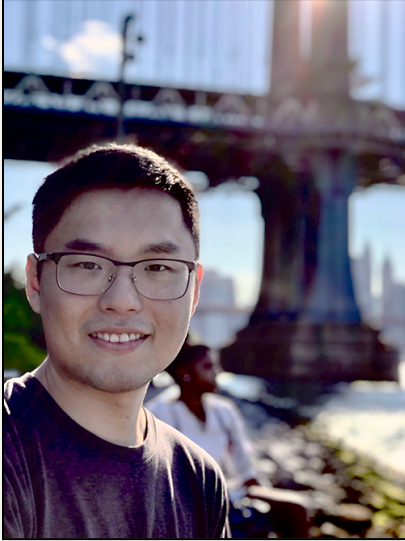
For Integrated Gradients (IG), we propose a novel non-uniform interpolation scheme to compute the IG attribution scores which replaces the baseline uniform interpolation. Our algorithm significantly reduces the total interpolation steps required without adversely impacting convergence. Experiments on the ImageNet



<p><b>Expected Graduation</b> <b>Date:</b> May 2025</p>	<p>dataset using a pre-trained InceptionV3 model demonstrate 2.6-3.6X performance speedup on GPU systems for iso-convergence. This includes the minimal 0.2-3.2% latency overhead introduced by the pre-processing stage of computing the non-uniform interpolation step-sizes.</p> <p>For GradCAM, we identify the root cause of the large computational overhead to be the dynamic run-time automatic differentiation. To overcome this issue, we propose to offload the gradient computation step to compile time via analytic evaluation. We validate the idea by designing an FPGA implementation of GradCAM that schedules the entire computation graph statically. For a TinyML ResNet18 model, we achieve a reduction in the explanation generation overhead from &gt;2X in software frameworks on CPU/GPU systems to &lt;0.01X on the FPGA using our designed hardware and static scheduling.</p> <p><b>Additional author(s):</b> Arijit Raychowdhury</p>
<p><b>2.16 Abhishek Moita</b> Yale University</p>  <p><b>Email:</b> abhishek.moitra@yale.edu <b>PI:</b> Priya Panda <b>Expected Graduation</b> <b>Date:</b> May 2024</p>	<p><b>Title:</b> XPer: Peripheral Circuit &amp; Neural Architecture Co-search for Area and Energy-efficient Xbar-based Computing</p> <p><b>Abstract:</b> The hardware efficiency and accuracy of Deep Neural Networks (DNNs) implemented on In-memory Computing (IMC) architectures primarily depend on the DNN architecture and the peripheral circuit parameters. It is therefore essential to holistically co-search the network and peripheral parameters to achieve optimal performance. To this end, we propose XPert, which co-searches network architecture in tandem with peripheral parameters such as the type and precision of analog-to-digital converters, crossbar column sharing, and layer-specific input precision using an optimization-based design space exploration. Compared to VGG16 baselines, XPert achieves 10.24× (4.7×) lower EDAP, 1.72× (1.62×) higher TOPS/W, 1.93× (3×) higher TOPS/mm<sup>2</sup> at 92.46% (56.7%) accuracy for CIFAR10 (TinyImagenet) datasets.</p> <p><b>Additional author(s):</b> Abhiroop Bhattacharjee, Youngeun Kim, and Priyadarshini Panda</p>



## 2.17 Tian Jin MIT



**Email:** tianjin@mit.edu

**PI:** Michael Carbin

**Expected Graduation**

**Date:** May 2026

**Title:** Pruning's Effect on Generalization Through the Lens of Training and Regularization

**Abstract:** Practitioners frequently observe that pruning improves model generalization. A long-standing hypothesis based on bias-variance trade-off attributes this generalization improvement to model size reduction. However, recent studies on over-parameterization characterize a new model size regime, in which larger models achieve better generalization. Pruning models in this over-parameterized regime leads to a contradiction -- while theory predicts that reducing model size harms generalization, pruning to a range of sparsities nonetheless improves it. Motivated by this contradiction, we re-examine pruning's effect on generalization empirically.

We show that size reduction cannot fully account for the generalization-improving effect of standard pruning algorithms. Instead, we find that pruning leads to better training at specific sparsities, improving the training loss over the dense model. We find that pruning also leads to additional regularization at other sparsities, reducing the accuracy degradation due to noisy examples over the dense model. Both effects are essential to explaining pruning's impact on generalization.

**Additional author(s):** Michael Carbin, Daniel M. Roy, Jonathan Frankle, Gintare Karolina Dziugaite



**2.18 Eric Atkinson**  
MIT



**Email:**

eatkinson@csail.mit.edu

**PI:** Michael Carbin

**Expected Graduation**

**Date:** September 2023

**Title:** A Language and Logic for Programming and Reasoning with Partial Observability

**Abstract:** Computer programs are increasingly deployed in partially-observable environments, whose state is not completely visible to the program but from which the program receives partial observations. Developers deal with partial observability by writing a state estimator that, given observations, deduces the hidden state of the environment. In safety-critical domains, to facilitate formally verifying safety properties, developers may write an environment model that captures the relationship between observations and hidden states.

In this work, we describe belief programming, a new programming methodology where developers write an environment model that the program runtime automatically uses to perform state estimation. To enable verification, we describe Epistemic Hoare Logic (EHL), which reasons about the possible states of belief programs the same way that classical Hoare logic reasons about classical programs. We demonstrate belief programming and EHL in a case study where we write and verify an engine descent controller for the Mars Polar Lander.

We show that size reduction cannot fully account for the generalization-improving effect of standard pruning algorithms. Instead, we find that pruning leads to better training at specific sparsities, improving the training loss over the dense model. We find that pruning also leads to additional regularization at other sparsities, reducing the accuracy degradation due to noisy examples over the dense model. Both effects are essential to explaining pruning's impact on generalization.

**Additional author(s):** Michael Carbin, Daniel M. Roy, Jonathan Frankle, Gintare Karolina Dziugaite



## POSTER SESSION 3

# Themes 3 and 4 Overview

Wednesday, May 17 @ 11:45 AM-1:00 PM

**Zones 3 and 4**

POSTER NO.	PRESENTER	TITLE
3.1	Akul Malhotra	TFix: Exploiting the Natural Redundancy of Ternary Neural Networks for Fault Tolerant In-Memory Vector Matrix Multiplication
3.2	Da Eun Shim	Vertically Integrated End-to-End Technology Evaluation Platform: BEOL and FEOL Co-design and Benchmarking
3.3	Imtiaz Ahmed	Utilizing Valley-Spin Hall Effect to enable Low Area and Energy Efficient XNOR-based In-Memory Dot Product Computations for Binary Neural Networks
3.4	Kartik Prabhu	MINOTAUR: Enabling Transformer Models at the Edge with Posits and Resistive RAM
3.5	Piyush Kumar	Vertically Integrated End-to-End Technology Evaluation Platform: MRAM-Interconnect Technology Co-design and Benchmarking
3.6	Siri Narla	Context-based Content Addressable Memory for Nearest Neighbor Search
3.7	Wangxin He	PRIVE: Efficient RRAM Programming with Chip Verification for RRAM-based In-Memory Computing Acceleration




3.8	Zhenyu Wang	HISIM: Heterogeneous Integration Simulator with 2.5D/3D Interconnect Modeling
4.9	Arghadip Das	Towards Energy-Efficient Collaborative Inference using Multi-System Approximations
4.10	Chris Richardson	Improving Commonsense in Dialogue through Self-Correction
4.11	Ruokai Yin	Multiplier-less Integer Quantization for Spiking Neural Networks
4.12	Soumendu Ghosh	Enabling Energy-Efficient Multimodal Transformers at the Edge
4.13	Tyler Lizzo	Tool Augmented LLaMA
4.14	Sakshi Choudhary	CoDeC: Communication-Efficient Decentralized Continual Learning
4.15	Sai Aparna Aketi	Neighborhood Gradient Clustering: An Efficient Decentralized Learning Method for Non-IID Data Distributions
4.16	Yuhang Li	DTSNN: Input-Aware Dynamic Timestep Spiking Neural Networks for Efficient In-Memory Computing
4.17	Zishen Wan	BERRY: Bit Error Robustness for Energy-Efficient Reinforcement Learning-Based Autonomous Systems
4.18	Guangyu Jiang	Exploring Collaborative Neuromorphic Swarms in Solving QUBO Problem



# Theme 3 Presentation Details

Wednesday, May 17 @ 11:45 AM-1:00 PM

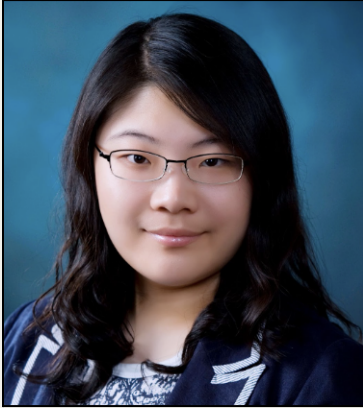
**Zone 3**

RESEARCH SCHOLAR	POSTER DETAILS
<p><b>3.1 Akul Malhotra</b> Purdue University</p>  <p><b>Email:</b> malhot23@purdue.edu <b>PI:</b> Sumeet Kumar Gupta <b>Expected Graduation Date:</b> December 2025</p>	<p><b>Title:</b> TFix: Exploiting the Natural Redundancy of Ternary Neural Networks for Fault Tolerant In-Memory Vector Matrix Multiplication</p> <p><b>Abstract:</b> In-memory computing (IMC) and quantization have emerged as promising techniques for edge-based deep neural network (DNN) accelerators by reducing their energy, latency, and storage requirements. In pursuit of ultra-low precision, ternary precision DNNs (TDNNs) hit the sweet spot in terms of achieving high efficiency without sacrificing much inference accuracy. In this work, we explore the impact of hard faults on IMC based TDNNs and propose TFix to enhance their fault tolerance. TFix exploits the natural redundancy present in most ternary IMC bitcells in conjunction with the high weight sparsity in TDNNs to provide up to 40.68% accuracy increase over the baseline with &lt; 6% energy overhead.</p> <p><b>Additional author(s):</b> Chunguang Wang, Sumeet Kumar Gupta</p>



### 3.2 Da Eun Shim

Georgia Tech



**Email:** daeun@gatech.edu

**PI:** Azad Naeemi

**Expected Graduation**

**Date:** August 2023

**Title:** Vertically Integrated End-to-End Technology Evaluation Platform: BEOL and FEOL Co-design and Benchmarking

**Abstract:** The goal of this task is to develop a vertically integrated end-to-end technology evaluation platform for various hardware motifs for cognitive systems based on CMOS and beyond CMOS devices. Toward that goal, we have been developing a PDK for 3nm GAAFET technology along with a range of interconnect technology options including those pursued in the SUPREME JUMP 2.0 Center. While our 3nm PDK is based on an open-source 3nm GAAFET developed by NCSU, we add our own assumptions on the number of nanosheets, BEOL dimensions, and interconnect resistances based on our TCAD models. We also create our own set of standard cells based on layouts and SPICE simulations for PnR of cognitive systems. In addition, we investigate promising interconnect-material candidates recommended through first-principles calculations and new descriptors of resistivity developed by the SUPREME center. Based on their recommendations, we aim to develop TCAD models of interconnects at the nm scale and test their compatibility and impact on cognitive systems.

**Additional author(s):** Piyush Kumar, Siri Narla

### 3.3 Imtiaz Ahmed

Purdue University




**Title:** Utilizing Valley-Spin Hall Effect to enable Low Area and Energy Efficient XNOR-based In-Memory Dot Product Computations for Binary Neural Networks

**Abstract:** Binary Neural Networks (BNNs) have shown immense promise for edge AI as their binarized synaptic weights and neuron activations can significantly reduce the compute, storage, and communication costs. Several works have explored XNOR-based BNNs using SRAMs and non-volatile memories. These designs typically need two bit-cells to encode signed weights, leading to an area overhead compared to conventional AND-based dot product computations. In this work, we address this limitation by exploring a compact and low-power In-Memory Computing (IMC) technique for XNOR-based dot products. Our approach utilizes Valley-Spin Hall





<p><b>Email:</b> ahmed202@purdue.edu <b>PI:</b> Sumeet Kumar Gupta <b>Expected Graduation Date:</b> December 2026</p>	<p>effect (VSH) in monolayer Tungsten di-selenide (<math>WSe_2</math>) to design bit-cells featuring access-transistor-less compact layout and differential encoding in a single device (which we exploit for XNOR-IMC). We co-optimize the VSH device and the memory arrays based on them to enable efficient in-memory dot product computations between signed binary inputs and signed binary weights. Our results show 4.8% ~ 9.0% and 36.6% ~ 62.5% lower compute latency and energy, respectively, with 49.3% ~ 64.4% smaller area compared to STT-MRAM and SOT-MRAM based XNOR-arrays.</p> <p><b>Additional author(s):</b> Karam Cho, Sumeet Kumar Gupta</p>
<p><b>3.4 Kartik Prabhu</b> Stanford University</p>  <p><b>Email:</b> kprabhu7@stanford.edu <b>PI:</b> Priyanka Raina <b>Expected Graduation Date:</b> June 2024</p>	<p><b>Title:</b> MINOTAUR: Enabling Transformer Models at the Edge with Posits and Resistive RAM</p> <p><b>Abstract:</b> Transformer models achieve state-of-the-art accuracy, but are challenging to run in resource-constrained edge environments, as they are large and difficult to quantize to 8b integers. MINOTAUR overcomes these challenges and enables both inference and training of machine learning models at the edge through (1) an alternative 8b floating-point data type, (2) a deep neural network accelerator optimized for operator fusion, and (3) temporal power-gating of on-chip non-volatile resistive RAM (RRAM). First, MINOTAUR uses 8b posits, an alternative to the IEEE-754 floating point standard that achieves a higher dynamic range for the same number of bits. Next, MINOTAUR efficiently accelerates Transformers with by fusing operations, enabled by a configurable vector datapath; this improves inference accuracy and reduces the number of memory accesses. Finally, MINOTAUR uses RRAM with software-controlled fine-grained temporal power gating to fit large models (e.g., 12 MB MobileBERT-tiny) entirely on-chip with minimal memory power.</p> <p><b>Additional author(s):</b> Jeffrey Yu</p>



### 3.5 Piyush Kumar

Georgia Tech



**Email:**

pkumar315@gatech.edu

**PI:** Azad Naeemi

**Expected Graduation**

**Date:** May 2024

**Title:** Vertically Integrated End-to-End Technology Evaluation Platform: MRAM-Interconnect Technology Co-design and Benchmarking

**Abstract:** The goal of this task is to develop a vertically integrated end-to-end technology evaluation platform for various hardware motifs for cognitive systems based on CMOS and beyond CMOS devices. Toward that goal, we have been working on a rigorous technology modeling framework for content addressable memories as the key building block for various cognitive systems such as hyperdimensional computing and neural networks. While SOT-based devices are promising for such applications due to their non-volatility and fast read/write speeds ( $<10\text{ns}$ ), there are major challenges in adopting them to the advanced technology nodes. In this work, we present a comprehensive framework for evaluating memory arrays while accounting for the impact of technology scaling and various back-end-of-line technologies. In collaboration with SUPREME research, various technology options are considered. Based on interconnect resistance values from TCAD simulations and MRAM device characteristics from experimentally validated/calibrated physical models, we quantify the potential array-level performance of MRAM using SPICE simulations.

**Additional author(s):** Da Eun Shim, Siri Narla

### 3.6 Siri Narla

Georgia Tech




**Email:** snarla6@gatech.edu

**Title:** Context-based Content Addressable Memory for Nearest Neighbor Search

**Abstract:** Spin Orbit Torque Magnetic random access memory (SOT-MRAM) are being considered for many unconventional circuit designs. Content addressable memories (CAMs) have proven themselves to be worthy contenders. CAMs have been used in many applications for performing similarity and nearest neighbor search (NNS) since they theoretically show very low  $O(1)$  complexity. However, challenges like minimum detectable Hamming distance, discharge current saturation, and energy and performance efficiency have limited their use. In this poster, we will present a novel



<p><b>PI:</b> Azad Naeemi <b>Expected Graduation Date:</b> May 2024</p>	<p>context-based CAM design using SOT-MRAMs that improves nearest neighbor search efficiency in CAM-based nearest neighbor applications. Using circuit-level data from SPICE simulations, recall rates for context-based CAM searches were studied on a Kaggle Car dataset with LSH encoding for various sparsity indices. Context-based CAMs can be useful in improving CAM applications like recommendation models, hyperdimensional computing, retrieval-based language models, etc. We also analyze the impact of technology scaling and various back-end-of-line technologies on CAMs.</p> <p><b>Additional author(s):</b> Piyush Kumar, Da Eun Shim</p>
<p><b>3.7 Wangxin He</b> ASU</p>  <p><b>Email:</b> <a href="mailto:wangxinh@asu.edu">wangxinh@asu.edu</a> <b>PI:</b> Jae-sun Seo <b>Expected Graduation Date:</b> August 2023</p>	<p><b>Title:</b> PRIVE: Efficient RRAM Programming with Chip Verification for RRAM-based In-Memory Computing Acceleration</p> <p><b>Abstract:</b> As deep neural networks (DNNs) have been successfully developed in many applications with continuously increasing complexity, the number of weights or parameters in DNNs surges, leading to consistent demands for denser memories than SRAMs. RRAM-based in-memory computing (IMC) achieves high density and energy efficiency for deep neural network inference, but RRAM programming remains to be a bottleneck due to high write latency and energy consumption. In this work, we present the Progressive-wRite In-memory program-VErify (PRIVE) scheme, which we verify with an RRAM test chip for IMC-based hardware acceleration for DNNs. We perform progressive write-verify on different bit positions of RRAM weights to enable error compensation and reduce programming latency/energy while still achieving high DNN accuracy. For 4-bit precision DNNs, PRIVE reduces the RRAM programming energy by 45%, while maintaining high accuracy of 91.91% (VGG-7) and 71.47% (ResNet-18) on CIFAR-10 and CIFAR-100 datasets, respectively.</p> <p><b>Additional author(s):</b> Jian Meng, Sujun Kumar Gonugondla, Shimeng Yu, Naresh R. Shanbhag and Jae-sun Seo</p>



## 3.8 Zhenyu Wang

ASU



**Email:** zwang586@asu.edu

**PI:** Yu (Kevin) Cao

**Expected Graduation**

**Date:** December 2024

**Title:** HISIM: Heterogeneous Integration Simulator with 2.5D/3D Interconnect Modeling

**Abstract:** Current monolithic design faces significant challenges of silicon area, fabrication cost, and data movement, especially when dealing with increasingly complex and diverse AI models. With advanced packaging, 2.5D and 3D interconnection today are able to provide high bandwidth and high channel density that are comparable to on-chip interconnect, inspiring new architectures and design paradigm. In this work, we propose HISIM, a benchmark tool for chiplet-based heterogeneous integration (HI), that evaluates the performance of monolithic, 2.5D, and 3D systems. HISIM emphasizes the hierarchical interconnection and associated data movement in the HI system. It integrates the technology roadmap of 2.5D/3D wires, converts the roadmap into electrical analysis, and performs a cycle-accurate simulation of the data flow. We apply HISIM to AI models, to analyze tradeoffs in terms of mapping methods. Our results highlight the advantages of 3D interconnect in performance and the emerging limitation of power consumption on algorithm mapping.

**Additional author(s):** Alper Goksoy, Umit Y Ogras, Chaitali Chakrabarti, Jae-sun Seo, Yu Cao



# Theme 4 Presentation Details

Wednesday, May 17 @ 11:45 AM-1:00 PM

**Zone 4**

RESEARCH SCHOLAR	POSTER DETAILS
<p><b>4.9 Arghadip Das</b> Purdue University</p>  <p><b>Email:</b> December 2027 <b>PI:</b> Vijay Raghunathan <b>Expected Graduation Date:</b> December 2026</p>	<p><b>Title:</b> Towards Energy-Efficient Collaborative Inference using Multi-System Approximations</p> <p><b>Abstract:</b> Collaborative inference applications based on distributed deep neural networks (DDNNs) are becoming increasingly popular. In these applications, DDNNs are used to classify 3D objects from a set of 2D images or views, known as multiview convolutional neural networks (MVCNN). However, due to the intensive computational demands, substantial communication overhead, high inference delay, and energy limits, it is difficult to deploy MVCNN on resource-constrained edge devices. We propose, for the first time, the concept of jointly optimizing distributed collaborative inference systems by employing a set of optimizations that explores the performance, accuracy, and energy trade-off space. Our proposed techniques prune the large design space using the non-uniform contribution of various perspectives/views in a multiview CNN to achieve an optimized quality-energy trade-off. In addition, we also propose a novel energy-aware heuristic, which dynamically configures edge inference systems based on application quality bounds and increases the system lifetime. Experimental results on an Intel Stratix IV FPGA development board-based prototype that executes 12-view 3D object classification show significant energy savings (2.6×–7.8×) for minimal (&lt;1%) application-level quality loss.</p> <p><b>Additional author(s):</b> Soumendu Kumar Ghosh, Arnab Raha, and Vijay Raghunathan</p>



## 4.10 Chris Richardson

Georgia Tech



**Email:**

crichardson8@gatech.edu

**PI:** Larry Heck

**Expected Graduation**

**Date:** May 2025

**Title:** Improving Commonsense in Dialogue through Self-Correction

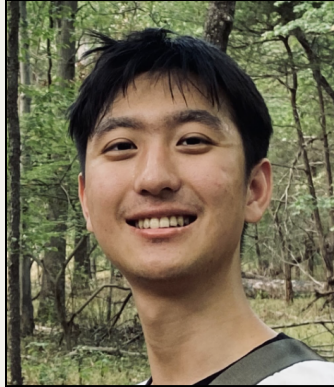
**Abstract:** Commonsense reasoning is a critical aspect of human communication that allows us to make inferences and draw conclusions based on our everyday experiences and understanding of the world. Despite rapid advances in conversational AI technology, driven by the advent of transformer architecture and large pre-trained language models, commonsense reasoning in conversational AI remains a challenging task. In this work, we propose a method for improving commonsense in dialogue response generation. To do so, we introduce \textsc{Syndicom} – a dataset of synthetic dialogues with commonsense, created from a knowledge graph and synthesized into natural language using GPT. The dataset includes valid and invalid responses to dialogue contexts, as well as human-written feedback for the invalid responses. We then propose a two-step procedure for improving dialogue responses, demonstrated on the invalid responses from \textsc{Syndicom}. Our procedure involves training a model to generate feedback for a dialogue response and a correction model that conditions on that feedback to improve a baseline response. Our method is scalable and involves no reinforcement learning. Empirical results demonstrate an improvement of our method over state-of-the-art chatbots like ChatGPT.

**Additional author(s):** Anirudh Sundar, Larry Heck



## 4.11 Ruokai Yin

Yale University



**Email:** ruokai.yin@yale.edu

**PI:** Priya Panda

**Expected Graduation**

**Date:** May 2026

**Title:** Multiplier-less Integer Quantization for Spiking Neural Networks

**Abstract:** We propose Multiplier-Integer (MINT) quantization, an efficient uniform quantization scheme for the weights and membrane potentials in spiking neural networks (SNNs). Unlike prior SNN quantization works, MINT quantizes the memory-hungry membrane potentials to extremely low bit-width (2-bit) to significantly reduce the total memory footprint. Additionally, MINT quantization shares the quantization scale between the weights and membrane potentials, eliminating the need for multipliers and floating arithmetic units, which are required by the standard uniform quantization. Experimental results demonstrate that our proposed method achieves accuracy that matches other state-of-the-art SNN quantization works while outperforming them on total memory footprint and hardware cost at deployment time. For instance, 2-bit MINT VGG-16 achieves 48.6% accuracy on TinyImageNet (+0.28% from the full-precision baseline) with approximately 93.8% reduction in total memory footprint from the full-precision model; meanwhile, our model reduces area by 93% and dynamic power by 98% compared to other SNN quantization counterparts.

**Additional author(s):** Yuhang Li, Abhishek Moitra, Priyadarshini Panda



## 4.12 Soumendu Ghosh

Purdue University



**Email:** ghosh37@purdue.edu

**PI:** Vijay Raghunathan

**Expected Graduation**

**Date:** July 2023

**Title:** Enabling Energy-Efficient Multimodal Transformers at the Edge

**Abstract:** In recent years, Multimodal AI (MMAI) has shown extensive promise to develop cognitive systems with a “deeper understanding of the world.” Recent studies have also proved the effectiveness of *transformers* in learning semantic video/audio/text representations as its design allows processing data of multiple modalities and, as a result, transformers form an integral building block of multimodal AI systems. However, the adoption of MMAI to real-world applications on resource-constrained edge platforms has been hindered by the intensive power, computation, and memory requirements of the underlying transformer workloads. To facilitate multimodal AI at the edge, we first explore performance and energy optimization techniques for each constituent subsystem such as multimodal sensors, communication, and compute in a complete IoT system running multimodal cognitive workloads. We also investigate inter-subsystem and inter-modal interactions and the impact of system-level quality-performance tradeoffs across sensing modalities. Based on the insights from these studies, we propose quality-scalable Multimodal Cognitive Systems to execute efficient multimodal inference at the edge.

**Additional author(s):** Arnab Raha, Arghadip Das, Vijay Raghunathan





#### 4.13 Tyler Lizzo

Georgia Tech



**Email:** lizzo@gatech.edu

**PI:** Larry Heck

**Expected Graduation**

**Date:** May 2028

**Title:** Tool Augmented LLaMA

**Abstract:** A central challenge to conversational AI is the balance between the size and accuracy of large language models (LLMs). Most prior research has not restricted the compute resources available to LLMs, allowing models on the order of hundreds of billions of parameters. For resource-constrained environments, techniques must be developed to bridge this computation gap while maintaining the system's accuracy. This project uses a neuro-symbolic approach, recently referred to as "augmented" LLMs. This approach provides a modular architecture that allows the integration of specialized downstream tools with the original pre-trained LLM. Given a set of downstream tools, our goal is to distill out the specialized knowledge of these tools while maintaining the LLM's general knowledge of these tools and open-domain conversational response generation. Our approach will potentially yield dramatic reductions in the model's size while maintaining accuracy. This project will utilize the open-source LLaMA models from Meta for implementation and evaluation.

**Additional author(s):** Larry Heck, Alwin Jin, Shrenik Bhansali

#### 4.14 Sakshi Choudhary

Purdue University



**Email:**

choudh23@purdue.edu

**PI:** Kaushik Roy


**Expected Graduation**

**Date:** May 2025

**Title:** CoDeC: Communication-Efficient Decentralized Continual Learning

**Abstract:** Privacy concerns prohibit the co-location of spatially as well as temporally distributed data at the edge, deeming it crucial to design training algorithms that enable efficient continual learning over decentralized private data. Decentralized learning allows serverless training with spatially distributed data. A fundamental barrier in such distributed learning is the high bandwidth cost of communicating model updates between agents. Moreover, existing works under this training paradigm are not inherently suitable for learning a temporal sequence of tasks while retaining the previously acquired knowledge. We propose CoDeC, a novel communication-efficient decentralized continual learning algorithm. We combine orthogonal gradient projection with gossip averaging across decentralized agents, alongside a novel lossless communication compression scheme. We

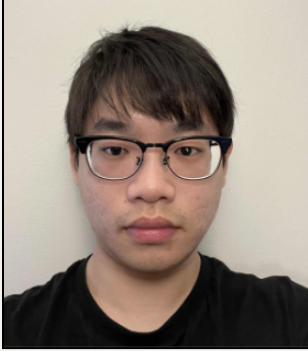


	<p>theoretically analyze the convergence rate for our algorithm and demonstrate through extensive experiments that CoDeC successfully learns distributed continual tasks with minimal forgetting. Our compression scheme results in up to 4.8x reduction in communication costs compared to the baseline.</p> <p><b>Additional author(s):</b> Sai Aparna Aketi, Gobinda Saha, Kaushik Roy</p>
<p><b>4.15 Sai Aparna Aketi</b> Purdue University</p>  <p><b>Email:</b> saketi@purdue.edu <b>PI:</b> Kaushik Roy <b>Expected Graduation Date:</b> December 2023</p>	<p><b>Title:</b> Neighborhood Gradient Clustering: An Efficient Decentralized Learning Method for Non-IID Data Distributions</p> <p><b>Abstract:</b> Decentralized learning algorithms enable the training of deep learning models over large distributed datasets, without the need for a central server. In practical scenarios, the distributed datasets can have significantly different data distributions across the agents. This work focuses on improving decentralized learning over non-IID data with minimal compute and memory overheads. We propose Neighborhood Gradient Clustering (NGC), a novel decentralized learning algorithm that modifies the local gradients of each agent using self- and cross-gradients. In particular, the proposed method replaces the local gradients with weighted mean of self-gradients, model-variant cross-gradients and data-variant cross-gradients. The data-variant cross-gradients are aggregated through an additional communication round without breaking the privacy constraints. Further, we present CompNGC, a compressed version of NGC that reduces the communication overhead by 32x through cross-gradient compression. We theoretically analyze the convergence characteristics of NGC and demonstrate its efficiency over non-IID data sampled from various vision and language datasets.</p> <p><b>Additional author(s):</b> Sangamesh Kodge, Kaushik Roy</p>



#### 4.16 Yuhang Li

Yale University



**Email:** yuhang.li@yale.edu

**PI:** Priya Panda

**Expected Graduation**

**Date:** May 2026

**Title:** DTSNN: Input-Aware Dynamic Timestep Spiking Neural Networks for Efficient In-Memory Computing

**Abstract:** Spiking Neural Networks (SNNs) have recently attracted widespread research interest as an efficient alternative to traditional Artificial Neural Networks (ANNs) because of their capability to process sparse and binary spike information and avoid expensive multiplication operations. Although the efficiency of SNNs can be realized on the In-Memory Computing (IMC) architecture, we show that the energy cost and latency of SNNs scale linearly with the number of timesteps used on IMC hardware. Therefore, in order to maximize the efficiency of SNNs, we propose input-aware Dynamic Timestep SNN (DT-SNN), a novel algorithmic solution to dynamically determine the number of timesteps during inference on an input-dependent basis. By calculating the entropy of the accumulated output after each timestep, we can compare it to a predefined threshold and decide if the information processed at the current timestep is sufficient for a confident prediction. We deploy DT-SNN on an IMC architecture and show that it incurs negligible computational overhead. We demonstrate that our method only uses 1.46 average timesteps to achieve the accuracy of a 4-timestep static SNN while reducing the energy-delay-product by 80%.

**Additional author(s):** Yuhang Li, Abhishek Moitra, Priyadarshini Panda



## 4.17 Zishen Wan

Georgia Tech



**Email:**

zishenwan@gatech.edu

**PI:** Arijit Raychowdhury

**Expected Graduation**

**Date:** July 2025

**Title:** BERRY: Bit Error Robustness for Energy-Efficient Reinforcement Learning-Based Autonomous Systems

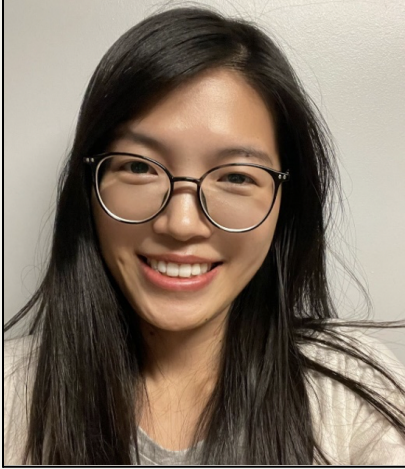
**Abstract:** Autonomous systems, such as Unmanned Aerial Vehicles (UAVs), are expected to run complex reinforcement learning (RL) models to execute fully autonomous position-navigation-time tasks within stringent onboard weight and power constraints. We observe that reducing onboard operating voltage can benefit the energy efficiency of both the computation and flight mission, however, it can also result in on-chip bit failures that are detrimental to mission safety and performance. To this end, we propose BERRY, a robust learning framework to improve bit error robustness and energy efficiency for RL-enable autonomous systems. BERRY supports robust learning, both offline and on-board the UAV, for the first time, and demonstrates the practicality of robust low-voltage operation on UAVs that leads to high energy savings in both compute-level operation and system-level quality-of-flight. We evaluate BERRY on 72 autonomous navigation scenarios and demonstrate that BERRY generalizes well across environments, UAVs, autonomy policies, operating voltages, and fault patterns, and consistently improves robustness, efficiency, and mission performance.

**Additional author(s):** Zishen Wan, Nandhini Chandramoorthy, Karthik Swaminathan, Pin-Yu Chen, Vijay Janapa Reddi, and Arijit Raychowdhury



## 4.18 Guangyu Jiang

Kennesaw State University



**Email:**

gjiang@students.kennesaw.edu

**PI:** Yan Fang

**Expected Graduation**

**Date:** July 2025

**Title:** Exploring Collaborative Neuromorphic Swarms in Solving QUBO Problem

**Abstract:** Combinatorial optimization problems prevail in engineering and industry. Some are NP-hard and thus become difficult to solve on edge devices due to limited power and computing resources. Quadratic Unconstrained Binary Optimization (QUBO) problem is a valuable emerging model that can formulate numerous combinatorial problems, such as Max-Cut, traveling salesman problems, and graphic coloring. QUBO model also reconciles with two emerging computation models, quantum computing and neuromorphic computing, which can potentially boost the speed and energy efficiency in solving combinatorial problems. In this work, we design a neuromorphic QUBO solver composed of a swarm of spiking neural networks (SNN) that conduct a population-based meta-heuristic search for solutions. The proposed model can achieve about x20~40 speedups on large QUBO problems in terms of time steps compared to a traditional neural network solver. As hardware codesigns for the proposal computing paradigm, we present our recent exploration in acceleration via the MAC operation primitive based on spintronic resonator arrays, and ICSRL's evaluation on a 40nm 25mW RRAM Compute-in-Memory Processor. Both demonstrate low latency and high energy efficiency in solving QUBO problems.

**Additional author(s):** Ashwin Sanjay Lele, Tariq Walker, Yan Fang



**CoCoSys**

CENTER FOR THE  
CO-DESIGN OF COGNITIVE SYSTEMS



**CoCoSys**

CENTER FOR THE  
CO-DESIGN OF COGNITIVE SYSTEMS