RESEARCH STATEMENT

Zhengqi Gao, MIT EECS

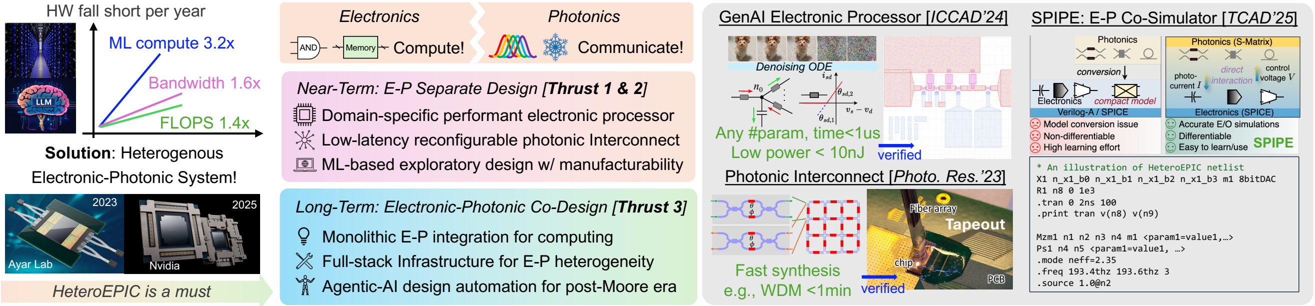
The explosive rise of generative AI, from trillion-parameter LLMs to vision diffusion models, has outpaced the scaling of existing electronic hardware and exposed fundamental limits in both computation and communication. Sustaining this rapid trajectory of AI requires rethinking the foundations of computing hardware itself. Within this landscape, I position heterogeneous electronic–photonic integrated circuits (**HeteroEPIC**s) as a cornerstone of next-generation AI hardware in the post-Moore era. HeteroEPIC integrates optical/photonic components with electronic counterparts either within the same package (i.e., co-packaged optics) or on the same substrate (i.e., monolithic integration), leveraging the complementary strengths of both domains — electronics excel at high-throughput, versatile computation, while photonics deliver high-bandwidth, low-latency communication.

The urgency of electronic–photonic co-integration is reaching an inflection point, exemplified by unprecedented investments from academia [1], industry (e.g., Ayar Labs TeraPHY 2023, Nvidia Spectrum-X 2025, Lightmatter L200 2025), and funding agencies (e.g., NAPMP, NSTC, and CARISSMA under the CHIPS Act, DARPA HAPPI 2024, and VLPI 2025). However, realizing the full vision of HeteroEPIC today is hindered by three systematic challenges: (i) breaking down disciplinary barriers between electronics and photonics, (ii) democratizing access to electronic–photonic system design, and (iii) ensuring scalability and robustness from prototype to manufacturing. In light of these challenges, my research focuses on **designing algorithms and hardware to advance heterogeneous electronic–photonic integrated circuits as the foundation of future AI systems.**

I am genuinely passionate and firmly believe that my work and research domain are **unique**, **distinguished**, and poised to make a revolutionary impact on the future of hardware for and beyond AI, because (i) my work embodies rare cross-domain expertise spanning electronics, photonics, and AI, demonstrated through more than 30 publications in top-tier venues (e.g., DAC, IEEE JLT, ICML) and prestigious awards in each area; (ii) my work combines rigorous theoretical contributions with an equal emphasis on experimental validation, including chip prototype implementations and on-chip verifications; and (iii) my research delivers immediate industrial value, while also tackling long-term fundamental scientific questions that will shape the field for years to come.

1 Past Research: AI-Driven Electronic-Photonic Design Automation

HeteroEPIC design can be conceptualized in two distinct phases: the present/near term and the long term. In the present and near term, electronics and photonics are typically fabricated independently and subsequently integrated at the packaging stage, with each domain developed in isolation as an independent sub-module. At this stage, design objectives capitalize on the intrinsic strengths of each domain: electronics target the development of domain-specialized processors optimized for peak performance on specific workloads (**Thrust 1**), while photonics focus on delivering robust, ultra–low-latency, and energy-efficient interconnect that mitigate electronic bandwidth limitations (**Thrust 2**). In the long term, realizing the full promise of heterogeneous integration demands genuine electronic–photonic co-design, enabling holistic, cross-layer optimization of both computation and communication to unlock the full system potential (**Thrust 3**). My prior research aligns directly with these objectives and is organized into three thrusts corresponding to these strategic directions. My overarching Ph.D. research was recognized with the First Place Award at the ACM/IEEE DAC Ph.D. Forum 2025.



```
* An illustration of HeteroEPIC netlist
X1 n_x1_b0 n_x1_b1 n_x1_b2 n_x1_b3 m1 8bitDAC
R1 n8 0 1e3
.tran 0 2ns 100
.print tran v(n8) v(n9)

Mzm1 n1 n2 n3 n4 m1 <param1=value1,…>
Ps1 n4 n5 <param1=value1, …>
.mode neff=2.35
.freq 193.4thz 193.6thz 3
.source 1.0@n2
```

Figure 1: Overview of background, my research vision, and three representative past works.

1.1 Thrust 1: AI-Driven Design for Domain Specialized Electronic Processors

GenAI Processor Hardware-Software Co-Design. Ordinary differential equation (ODE)-driven neural networks, such as Neural ODEs, diffusion models, and state-space models (SSMs), have become cornerstone technologies in generative AI. However, their computation requires solving ODEs, which introduces inherent temporal dependence $x_{t+1} = x_t + \Delta t \cdot f_\theta(x_t)$ that cannot be paralleled on existing processors such as GPUs. To overcome this bottleneck, we first propose KirchhoffNet, a specialized hardware accelerator for generative AI [2]. Our key insight is that the temporal variable t in these models is a unitless numerical parameter for integration purpose; by physically realizing an ODE in an electronic integrated circuit using Kirchhoff Current Law (KCL), we bestow t a physical unit. Adjusting the on-chip capacitances tunes the RC time constants, enabling substantial acceleration (see Figure 1 for details). Mapping a given diffusion model onto the KirchhoffNet hardware is done by knowledge distillation inspired by our previous work spotlighted at ICLR'22 (top 5%) [3]. The on-chip training is done by our proposed adjoint method [4], which has been adopted in UCSD CSE 291E circuit course in Fall 2024. Our results show that by scaling the capacitors in the circuit, KirchhoffNet can execute diffusion models $<1\,\mu$s, consuming only about 10 nJ of energy with a chip area footprint of several hundred mm^2. In collaboration with another MIT team, we have demonstrated a CMOS-based prototype as shown in Figure 1. This work has attracted strong interest from academia (e.g., UCSD, Duke, CUHK) and industry (e.g., Lam Research, Apple, Rain AI), and was recognized with the prestigious 2024 MLSys Rising Star Award by MLCommons.

Agile and Manufacturable HW Design via Machine Learning. Modern processor design demands deep domain expertise, long and costly iteration cycles, and significant manual intervention. I have pioneered machine learning approaches that dramatically accelerate the transition from concept to fabricated silicon. For analog circuits (e.g., Opamp for processor I/O, PLL for clock generation), we introduce a multi-objective Bayesian optimization framework leveraging Bayesian neural networks [5], which reduces simulation time by 2× compared to existing methods at the time, while delivering outstanding design quality. This work has inspired ∼40 follow-up studies within a few years. For the digital domain (e.g., adder, multiplier), I recognized that the true bottleneck lies in front-end HDL generation, dominated by manual coding and prone to human error. We have addressed this critical bottleneck with a token entropy-aware reinforcement learning (RL) framework that guides LLMs for Verilog generation [6], outperforming existing standard RL (e.g., GRPO, PPO) and supervised finetuning baselines by absolute 5–10% on VerilogEval and RTLLM. We further extend this approach with a co-evolving verifier LLM, achieving state-of-the-art performance on general LLM reasoning benchmarks [7] at 7B/8B scale.

Beyond automation, rapidly diagnosing processor performance under process variations is critical at sub-3nm nodes, where increased manufacturing variability and a growing number of process, voltage, and temperature (PVT) corners threaten yield and reliability. We have spearheaded a series of Bayesian inference techniques for parametric yield estimation [8–10] across multiple corners and for predicting extremely low failure rates in processor memory. For instance, our NOFIS method [10] leverages a normalizing flow–based proposal distribution to guide importance sampling, reducing the number of simulation times by 2∼3× over baseline approaches while maintaining the same estimation accuracy. This work was recognized with the MIT MARC 2024 Best Pitch Award.

1.2 Thrust 2: AI-Driven Full-Stack Design Automation for Photonic Interconnect

System Foundation: Intelligent Synthesis and Routing for Programmable Photonics. Integrated programmable photonics has recently emerged as a transformative platform for reconfigurable photonic interconnect [11]. Unlike conventional fixed-topology optical interconnect such as Benes or Clos networks, recirculating programmable photonics can implement infinite impulse response (IIR) filters, creating a unique opportunity to embed signal processing directly within communication pathways. I realized that the functional synthesis and routing capabilities of such hardware had not been systematically explored and set out to address this gap, aiming to remove a key barrier to its wide adoption. As a first step, we analytically derive the system-level transfer function to assess whether closed-form synthesis is feasible using mathematical induction [12]. This work was recognized as an Editor's Highlight in Optica Photonic Research 2023 because the transfer function had not been known before our contribution. Building on this foundation, we develop a differentiable simulator and a gradient-based synthesis approach that supports arbitrary frequency-domain optical functions and diverse recirculating topologies, achieving synthesis in under one minute, a 100× improvement over manual trial-and-error [13]. The effectiveness of this approach has been validated on real hardware tapeouts (Figure 1), and a similar method has since been adopted by IPRONICS, a leading commercial programmable photonic startup. We also address the routability of programmable photonics, mathematically proving key characteristics such as the maximum, minimum, mean, and standard deviation of routable optical interconnect [14].

Physical Foundation: Learning-Based Swift and Resilient Photonic Device Design. Building photonic interconnect relies on fundamental silicon photonic devices such as Mach–Zehnder modulators (MZMs), waveguides, and couplers. However, these devices are highly sensitive to manufacturing variations. I have advanced a set of modeling and design techniques that enable fast, variation-aware photonic device design. First, we leverage neural operators [15] to rapidly estimate electromagnetic fields in photonic devices, achieving a 100× speedup over traditional numerical solvers and working as a surrogate simulator for swift design space exploration. Next, we propose an automatic synthesis approach for key photonic devices based on Bayesian optimization [16], combined with a Pareto-front (PF) selection criterion to ensure robustness [17]. This framework generates high-quality designs within approximately 100 simulations, comparable to or fewer than state-of-the-art adjoint inverse design methods, while remaining gradient-free, significantly simplifying implementation and deployment. Our methods have attracted strong industrial interest: Ansys Lumerical has engaged with us for integration into their product suite, and our PF selection approach was recognized as an Editor's Highlight by Optica Express 2024.

1.3 Thrust 3: Electronic-Photonic Co-Design and HeteroEPIC Analysis

The trend in HeteroEPIC is to place photonic components ever closer to their electronic counterparts, in some cases to within less than 1 µm on the same substrate. At such proximity, the two domains can no longer be treated as isolated; their tight coupling requires accurate co-simulation and joint optimization. A unified co-simulation framework is essential for this goal, yet remains absent. To address this, we have developed SPIPE [18], the first electronic–photonic co-simulator that preserves native formalisms for both domains, S-matrix for photonics and SPICE for electronics (Figure 1). SPIPE matches the accuracy of analytical solutions and commercial tools such as Lumerical INTERCONNECT while delivering 2∼85× speedup. Importantly, SPIPE enables, for the first time, the computation of derivatives of electronic signals (voltage/current) with respect to photonic signals, unlocking gradient-based cross-domain joint optimization. SPIPE has attracted many inquires from industry and academia (e.g., Cadence, and Texas A&M University). Additionally, we have developed LightSim, an error-aware behavioral simulator for accurately emulating large-scale AI workloads on HeteroEPIC. LightSim captures device-level imperfections and quantifies their impact on application-level performance for workloads such as diffusion models and LLMs. Our experiments uncovered critical design guidelines not visible to existing tools—for example, the need for wavelength channel spacing below 0.5 nm in MZM meshes to prevent severe LLM hallucinations, and the fundamental trade-off between photonic tensor-core size and numerical precision with 8 bits as a limit. These results establish LightSim as a game-changing tool for guiding architectural decisions in next-generation AI on HeteroEPIC, and are detailed in a forthcoming manuscript.

As a next step, we are building the system-level counterpart to SPIPE: the first Verilog-like HDL for HeteroEPIC, enabling unified simulation, synthesis, and layout across electronic and photonic components. This will complete the foundation for a full-stack design automation flow that seamlessly spans both electronic and photonic domains, which has never before realized in HeteroEPIC design.

2 Future Directions: AI and Automation (A^2) Lab

My future research will build on my prior work to establish HeteroEPIC as a foundational hardware paradigm and to advance its full-stack ecosystem, spanning algorithms to hardware. In the long term, I aim to extend to broader post-Moore computing systems. Specifically, I will focus on the following core aspects, presented in order **from current priorities to long-horizon efforts**: building foundations (§2.1), broadening accessibility (§2.2), maximizing yield (§2.3), and ultimately providing full life-cycle support for real-world deployment (§2.4).

2.1. Foundational Infrastructure for Very Large-Scale HeteroEPIC. Realizing HeteroEPIC at scale demands transformative infrastructure innovations, an ambition already drawing strong interest from initiatives such as DARPA VLPI. I will elevate my prior co-simulator SPIPE [18] into a scalable, widely adopted platform that serves as the backbone for both academic research and industrial deployment. In addition, I will drive the creation of a comprehensive ecosystem by: (i) developing placement and layout tools for monolithic and chiplet-based electronic-photonic systems; (ii) designing Verilog-augmented HDL tailored for HeteroEPIC; (iii) creating unified electronic-to-photonic end-to-end optimization frameworks for co-packaged optics and advanced 3D integration; (iv) building verification and sign-off methodologies that jointly consider electrical, optical, and thermal constraints; (v) enabling interoperability through open data formats and APIs for heterogeneous design flows; (vi) developing architectural tools to standardize power, performance, and area analysis for HeteroEPIC within the community; and (vii) collaborating with circuit design groups to fabricate large-scale HeteroEPIC using the proposed tools.

2.2. Agentic-AI Design Automation for HeteroEPIC and Beyond. Integrated circuit design flows are inherently text-centric (e.g., SPICE netlists, HDL code, and GDS layout files), making agent-based approaches a natural fit for selected stages of the workflow. Building on my prior work in applying RL for LLM reasoning [7] and

Verilog generation [6], I plan to extend the use of LLMs to a broader range of tasks within the design flow of HeteroEPIC. Future directions include, for example: (i) developing a fully open-source HeteroEPIC design agent built on OpenROAD and Model Context Protocol (MCP); (ii) enabling cross-stack HDL generation that incorporates physical design metrics such as power, area, and timing; (iii) advancing LLM-assisted analog electronic or photonic circuit optimization beyond conventional transistor sizing; (iv) and creating ReAct-based vision-language models capable of interpreting testbench visual results and generating actionable feedback. Additionally, these methods will be specialized to emerging hardware paradigms that are related to, yet extend beyond, HeteroEPIC, such as neuromorphic photonics, optoelectronics for quantum computing.

2.3. Design-for-Manufacturing HeteroEPIC in the CHIPS Era. Achieving high yield is central to the viability of HeteroEPIC and the semiconductor industry at large. In practice, design for manufacturing (DFM) provides the foundation for yield maximization by bridging design constraints with process variability. Within my broader body of work, a distinct line has addressed DFM in HeteroEPIC, demonstrating that process, material, device, or circuit properties often must be inferred from a few measurements [8, 19, 20]. Building on the national momentum to restore semiconductor manufacturing leadership under the CHIPS Act, my future research seeks to overcome these limitations as a critical step toward revitalizing and advancing both HeteroEPIC manufacturing and semiconductor design more broadly. Specifically, my future research will focus on pushing yield upward by: (i) embedding physics-based priors into ML models for process, device, circuit, and system-level manufacturing analysis; (ii) creating few-shot anomaly detection techniques for time-series data representing manufacturing, operational, or reliability signals [21]; (iii) enabling privacy-preserving, multi-institutional hardware and manufacturing modeling through federated learning; and (iv) advancing algorithms for efficient design space exploration of semiconductor processes, devices, and systems.

2.4. Operational-Cycle Support through Hardware–Software Co-Design. Beyond the design cycle, realizing the long-term impact of HeteroEPIC requires full life-cycle support at runtime and application levels. I envision HeteroEPIC as enabling computation directly during signal propagation through photonic interconnects, a paradigm that departs fundamentally from existing software stacks designed solely for processor execution and that largely overlook opportunities along communication pathways. Supporting this vision will demand hardware–software co-design tightly coupled with accurate cross-domain modeling and workload-driven optimization. Building on my past work in functional synthesis for photonic interconnects [12], I will investigate how to architect HeteroEPIC and other emerging heterogeneous hardware to sustain real-world deployment across diverse applications. Example directions include: (i) enabling in-transit signal computation within photonic interconnects to reduce data movement overhead, latency, and energy consumption; (ii) designing adaptive scheduling and resource allocation for multi-tenant, latency-critical workloads on edge and embedded systems; (iii) creating unified co-optimization strategies that jointly refine model architectures, quantization schemes, and hardware mappings to maximize performance-per-watt; and (iv) developing high-fidelity simulation and benchmarking frameworks to evaluate trade-offs in performance, energy, and quality of service for next-generation workloads.

3 Collaboration Plans and Funding Opportunities

My research is inherently interdisciplinary, from low-level physics, to circuit and system, and up to high-level algorithms and deep learning. This unique breadth of expertise enables me to identify pressing research challenges at an early stage and to lead innovations in the post-Moore era. In large, multi-investigator, center-scale efforts, I can play the role of connecting experts across materials, hardware, and algorithms, enabling cohesive collaboration.

In my future career, I plan to collaborate broadly across the EECS and materials communities, with a particular emphasis on partnerships that advance my research vision. These include but are not limited to: (i) engaging with the EDA community to develop advanced automation frameworks for HeteroEPIC design; (ii) partnering with computer architecture groups for systematic simulation and performance evaluation of HeteroEPIC; (iii) collaborating with electronic and photonic circuit design experts to fabricate prototypes and validate my design automation tools; and (iv) working with ML researchers to explore AI-driven approaches for IC design.

Potential funding sources include federal agencies such as NAPMP, NSTC, and CARISSMA under the CHIPS Act, NSF (e.g., ECCS, CISE), AFOSR, ONR, and DARPA, as well as industrial partners like Google, Meta, Cadence, Lam Research, Analog Devices, Samsung, Sandia National Laboratories (SNL), and TSMC. I have already established connections with several of these organizations.

References

[1] Amir H Atabaki, Sajjad Moazeni, Fabio Pavanello, Hayk Gevorgyan, Jelena Notaros, Luca Alloatti, Mark T Wade, Chen Sun, Seth A Kruger, Huaiyu Meng, et al. Integrating Photonics with Silicon Nanoelectronics for the Next Generation of Systems on a Chip. *Nature*, 2018.

[2] **Zhengqi Gao**, Fan-Keng Sun, Ron Rohrer, and Duane S Boning. KirchhoffNet: A Scalable Ultra-Fast Analog Neural Network. In *IEEE/ACM ICCAD*, 2024.

[3] Zihui Xue*, **Zhengqi Gao***, Sucheng Ren*, and Hang Zhao. The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation. In *ICLR*, 2022. [*Equal contribution].

[4] Jiahua Li*, Danyal Ahsanullah*, **Zhengqi Gao***, and Ron Rohrer. Circuit Theory of Time Domain Adjoint Sensitivity. *IEEE TCAD*, 2023. [*Equal contribution].

[5] **Zhengqi Gao**, Jun Tao, Fan Yang, Yangfeng Su, Dian Zhou, and Xuan Zeng. Efficient Performance Trade-off Modeling for Analog Circuit based on Bayesian Neural Network. In *IEEE/ACM ICCAD*, 2019.

[6] Jiahe Shi, **Zhengqi Gao**, Ching-Yun Ko, and Duane S. Boning. Verilog Code Generation with Large Language Models. *DATE*, 2026. [Under review, title anonymized].

[7] Kaiwen Zha*, **Zhengqi Gao***, Maohao Shen, Zhang-Wei Hong, Duane S. Boning, and Dina Katabi. RL Tango: Reinforcing Generator and Verifier Together for Language Reasoning. In *NeurIPS*, 2025. [*Equal contribution].

[8] **Zhengqi Gao**, Jun Tao, Yangfeng Su, Dian Zhou, Xuan Zeng, and Xin Li. Efficient Rare Failure Analysis Over Multiple Corners via Correlated Bayesian Inference. *IEEE TCAD*, 2020.

[9] **Zhengqi Gao**, Jun Tao, Dian Zhou, and Xuan Zeng. Efficient Parametric Yield Estimation Over Multiple Process Corners via Bayesian Inference Based on Bernoulli Distribution. *IEEE TCAD*, 2020.

[10] **Zhengqi Gao**, Dinghuai Zhang, Luca Daniel, and Duane S Boning. NOFIS: Normalizing Flow for Rare Circuit Failure Analysis. In *DAC*, 2024.

[11] Wim Bogaerts, Daniel Pérez, José Capmany, David AB Miller, Joyce Poon, Dirk Englund, Francesco Morichetti, and Andrea Melloni. Programmable Photonic Circuits. *Nature*, 2020.

[12] **Zhengqi Gao**, Xiangfeng Chen, Zhengxing Zhang, Uttara Chakraborty, Wim Bogaerts, and Duane S. Boning. Automatic Synthesis of Light-Processing Functions for Programmable Photonics: Theory and Realization. *Photon. Res.*, 2023.

[13] **Zhengqi Gao**, Xiangfeng Chen, Zhengxing Zhang, Uttara Chakraborty, Wim Bogaerts, and Duane S. Boning. Gradient-Based Power Efficient Functional Synthesis for Programmable Photonic Circuits. *Journal of Lightwave Technology*, 2024.

[14] **Zhengqi Gao**, Xiangfeng Chen, Zhengxing Zhang, Chih-Yu Lai, Uttara Chakraborty, Wim Bogaerts, and Duane S. Boning. Provable Routing Analysis of Programmable Photonic Circuits. *Journal of Lightwave Technology*, 2024.

[15] Jiaqi Gu, **Zhengqi Gao**, Chenghao Feng, Hanqing Zhu, Ray Chen, Duane S. Boning, and David Pan. Neurolight: A Physics-Agnostic Neural Operator Enabling Parametric Photonic Device Simulation. In *NeurIPS*, 2022.

[16] **Zhengqi Gao**, Zhengxing Zhang, and Duane S. Boning. Automatic Synthesis of Broadband Silicon Photonic Devices via Bayesian Optimization. *JLT*, 2022.

[17] **Zhengqi Gao**, Zhengxing Zhang, Zichang He, Jiaqi Gu, David Z. Pan, and Duane S. Boning. Selecting Robust Silicon Photonic Designs after Bayesian Optimization without Extra Simulations. *Opt. Express*, 2024.

[18] **Zhengqi Gao**, Jiaqi Gu, Luca Daniel, Ron Rohrer, and Duane S. Boning. SPIPE: Differentiable SPICE-Level Co-Simulation Program for Integrated Photonics and Electronics. *IEEE TCAD*, 2025.

[19] **Zhengqi Gao** and Ron Rohrer. Efficient Non-Monte-Carlo Yield Estimation. *IEEE TCAD*, 2022.

[20] Zhengxing Zhang, Milica Notaros, **Zhengqi Gao**, Uttara Chakraborty, Jelena Notaros, and Duane S. Boning. Impact of Process Variations on Splitter-Tree-Based Integrated Optical Phased Arrays. *Opt. Express*, 2023.

[21] Chih-Yu Lai, Fan-Keng Sun, **Zhengqi Gao**, Jeffrey H Lang, and Duane Boning. Nominality Score Conditioned Time Series Anomaly Detection by Point/Sequential Reconstruction. In *NeurIPS*, 2023.