Nippun Sabharwal

+14479023085 | nippuns2@illinois.edu | linkedin.com/in/nippun-sabharwal | github.com/nippun-live

Education

University of Illinois Urbana-Champaign *BS in Computer Engineering* **Aug 2022 - May 2026**
Honors: Samsung Engineering Scholarship, Illinois Outstanding Scholarship, Illinois Engg Achievement Scholarship
Coursework: Deep Learning, Parallel Programming, Computer Architecture, Data Structures, Algorithms

Skills

Python, Java, C, C++, CUDA, SQL, PyTorch, TensorFlow, XGBoost, OpenCV, LangChain, HuggingFace, Linux, Git, ETL, AWS (Bedrock, SageMaker, EC2), Azure(Core, Compute), Docker, Kubernetes, Postgres, MongoDB, Tableau,

Work Experience

AstraZeneca **May 2025 - August 2025**
Machine Learning Intern Gaithersburg, Maryland

- **Developed an intelligent agent** with a natural language interface using **RAG and LangGraph with the MCP** framework that increased data accessibility by querying millions of rows of data under 10 s
- Designed a secure, highly available **AWS stack** (ECS, Application Load Balancer, Secrets Manager, Web Application Firewall, CloudWatch alerting) that delivers 99.9 percent measured uptime and scales automatically to 50+ concurrent users.
- Led **cross-functional discovery**, built the minimum viable product independently, then collaborated with the Mexico Data Science team for **company-wide rollout**.

Stealth Startup, Robotics **August 2025**
Intern, World Models & Hardware San Francisco, CA

- **Spearheaded performance tuning of world model**, developing mixed precision CUDA kernels for latent autoregressive **flow-matching architecture**, raising H100 training throughput 2.3x and cutting GPU memory use.
- **Set up hardware and drivers for the Unitree G1 humanoid** , integrating encoders, IMUs, force-torque sensors, and RealSense depth into a DORA pipeline designed for sub-25 ms latency at 250 Hz for **teleoperation**
- **Developing URDFs and a low-level motor control stack** in C++ with a whole-body controller, enabling stable standing and dynamic walking within two weeks of hardware arrival

Mashreq Bank **May 2023 - Jul 2023**
Cloud Infrastructure Intern

- Developed and optimized **RESTful & GraphQL APIs for** customer-facing applications, improving data retrieval speed by 40% and scalability for high-traffic usage.
- Refined CI/CD pipelines using **Jenkins and Docker**, reducing deployment time by 50%, which allowed for faster iterations in production environments.

Research

(NeurIPS 2025) Toward Engineering AGI: Benchmarking the Engineering Design Capabilities of LLMs
Guo, X., Li, Y., Kong, X., Sabharwal, N., et al. (2025). Proceedings of the 39th Conference on Neural Information Processing Systems

- Pioneered **systematic evaluation framework** for assessing frontier LLMs on complex **hardware design synthesis**, benchmarking models across hierarchical SystemVerilog tasks (sequential logic, memory structures, encoding schemes)
- Architected **closed-loop agentic system** with automated verification pipeline integrating simulation and synthesis tools, enabling iterative refinement through structured error feedback, achieved synthesizable designs through autonomous correction cycles

National Center for Supercomputing Applications, Computer Vision Researcher **Jan 2025 - Present**

- Leading research in **state-of-the-art video restoration techniques** using unsupervised, physics-informed deep learning frameworks for modeling and simulating atmospheric turbulence and fluid dynamics.
- Integrated **Vision Transformers, NeRFs, and 3D Mamba** based architectures with **optical-flow** estimation and **3D geometric scene reconstruction** to accurately predict and correct non-rigid distortions in real-world videos.

Projects

RISC-V Operating System Kernel and IO Drivers (https://github.com/nippun-live/riscvos)

- **Built a UNIX-style kernel** in C with bootloader, trap handling, Sv39 demand paging, cooperative and preemptive threading, sustaining stable multitasking in QEMU.
- **Designed a write-ahead-logged, block-based filesystem** with a configurable **write-back cache**, cutting IO latency by roughly 45 percent while guaranteeing crash-safe metadata consistency across VirtIO and in-memory devices.

Robotic Teleoperation with VR (Demo: https://nippun-live.github.io/)

- Built WallE, a VR teleoperation interface for intuitive robot control over distances up to 3,000 miles.
- Trained imitation learning policies for long-horizon tasks like sorting, insertion, and folding.
- Integrated 3-DOF Oculus head tracking with an IK robotic neck, cutting fine manipulation time by about 30 percent.