Yuchen You

Unit 5030, 1929 Plymouth Rd., Ann Arbor, MI 48105

Email : yuchenxr@umich.edu
Alt : wesley_you@sjtu.edu.cn
Mobile : (+1) 734 510 0456

Education

- **University of Michigan** 2024 – Present
  B.S.E in Computer Science - GPA 3.94/4.00 *Ann Arbor, MI, US*
  - **Honors**: Cheng-Family Scholarship (Top 2%)
- **Shanghai Jiao Tong University** 2022 – 2024
  B.S.E in Mechanical Engineering - GPA 3.83/4.00 *Shanghai, China*
  - **Honors**: John Wu & Jane Sun Sunshine Scholarship (Top 6%), SJTU Excellence Scholarship (Top 3%)

Research Experience

- **Agentic Distributed System Ops** May 2024 – Present
  Order Lab, University of Michigan; Advisor: Prof. Ryan (Peng) Huang *Ann Arbor, MI, US*
  - **Closed-loop distributed control plane**: Engineered a **closed-loop distributed control plane** for **ZooKeeper** clusters to mitigate common distributed failures like overload failure and network fluctuation; integrated **Prometheus** for telemetry, **HAProxy** for traffic shaping, and **Resilience4j** for circuit breaking.
  - **Secure execution substrate**: Developed a **secure execution substrate** by encapsulating mitigation logic—spanning network throttling (**tcconfig**), dynamic load balancing (**HAProxy Runtime API**), **JVM** memory tuning, and disk I/O control (**fsync**)—into a library of **atomic, verified bash scripts**, ensuring deterministic and safe recovery actions regardless of the controller type.
  - **Load injection framework**: Engineered a custom **load injection framework** (zkbench) to reconstruct real-world failure patterns: supported **multi-staged traffic lifecycles** (warmup→spike→cooldown) and **dynamic skewed distributions** (weighted/random) to stress-test cluster resilience under unbalanced pressure.
  - **Static baseline & evaluator**: Developed a **fine-tuned static control plane** as a high-performance baseline: integrated a **YAML-configured detector** with the **atomic bash library** for deterministic mitigation, and built a **closed-loop evaluator** that correlates injection rates with system throughput to quantify precise recovery fidelity.
  - **Agentic reasoning layer**: Designed a **general-purpose Agentic reasoning layer** (with GPT-4o model) implementing the **Model Context Protocol (MCP)** to standardize tool interfaces; utilized **Chain-of-Thought (CoT)** to synthesize **high-efficiency analysis** of global system states from raw telemetry, adaptively orchestrating atomic actions to outperform the static baseline.
- **SoftRobot Electronic Control** Sept. 2024 – Present
  HDR Lab, University of Michigan; Advisor: Prof. Xiaonan (Sean) Huang *Ann Arbor, MI, US*
  - **Heterogeneous real-time control stack**: Architected a **heterogeneous real-time control stack** spanning **STM32** (actuation) and **Orange Pi** (planning); designed custom PCBs and implemented **multi-threaded task isolation** to decouple high-frequency **CAN/I^{2}C** telemetry from computational logic.
  - **Hybrid control strategy**: Developed a hybrid control strategy combining classic **PID** feedback with data-driven **Model Predictive Control (MPC)**; integrated ResNet (with PyTorch) for neural networks to optimize system parameters online, enhancing tracking accuracy under non-linear soft-robot dynamics.
  - **Kinematic solver**: Engineered a high-performance kinematic solver based on **Piecewise Constant Curvature (PCC)** and dynamic modeling; implemented optimized **Jacobian Inverse Kinematics (IK)** algorithms to resolve high-DoF redundancy in real-time, utilizing **convex optimization** to minimize configuration energy.
  - **Validation and recognition**: Validated system performance through industry collaboration with **General Motors**; work recognized with an **ICRA 2025 Best Poster Award** (Atlanta) and accepted spotlight talks at **RoboSoft 2025** and **ICON 2025**.
- **Control Developer** Feb. 2024 – Sept. 2024
  Sirius Lab, Shanghai Jiao Tong University; Advisor: Prof. Yutong Ban *Shanghai, China*
  - **Embodied AI pipeline**: Engineered an **Embodied AI pipeline** integrating **LLMs** for high-level task planning and a **ZED Depth Camera** for 3D perception, enabling a **Flexiv 7-DoF** arm to execute complex reasoning tasks (e.g., jigsaw puzzles) from natural language instructions.
  - **Motion control**: Lead the development of precise motion control algorithms based on the **Flexiv-RDK**; implemented **Inverse Kinematics (IK)** solvers and collision-free trajectory planning to translate abstract LLM plans into executable, smooth motor commands.
  - **Sim-to-Real**: Bridged the **Sim-to-Real gap** by incorporating simulation data into the control flow, refining path planning robustness in virtual environments before deploying on physical hardware.

Publications

- **Origami Inspired Soft Robotic Arm: A Modular Platform for Manipulation** May 2025
  IEEE ICRA Workshop, Atlanta, USA *Best Poster Award*
  - **Authors**: Jiyang Wang, **Yuchen You**, Xinqi Zhang, Haobo Fang, Jiaqi Wang, Xiaonan Huang

Patent

- **U.S. Patent, 2025**: J. Wang, Y. You, X. Xhang, H. Fang, J. Wang, and X. Huang, "Lightweight, Proprioceptive, Origami-Inspired Soft Robotic Arm for High Payload, Low-Cost Reconfigurable Manipulation," U.S Patent, (Pending), 2025.

Selected Projects

- **CUDA Proxy Player** Aug. 2025 – Dec. 2025
  CSE 582 (Advanced Operating Systems) – CUDA runtime optimization for LLM inference *University of Michigan*
  - **Multi-path CUDA execution proxy**: Designed a multi-path CUDA execution proxy (Baseline / Worker / Graph) with automatic routing based on batch size and micro-op count to cut CPU launch overhead for LLM inference workloads.
  - **Graph capture, caching, and bucketing**: Implemented CUDA Graph capture, caching, and update in a centralized graph manager; added shape bucketing (64–4096, 64-aligned) and per-bucket memory pools, achieving **1.5–3**× speedup over the baseline on repetitive workloads.
  - **Persistent workers and device queue**: Built persistent worker kernels (grid-stride loops) with a lock-free device buffer, reducing micro-op launch overhead and queue latency to achieve speedup for small-batch, many-op scenarios.
  - **Micro-op coverage and cuBLAS integration**: Extended micro-op kernels (BiasAdd, GELU, ReLU, LayerNorm, KV-cache, Embedding) and integrated cuBLAS GEMM into the proxy to cover mainstream LLM inference paths.
  - **Instrumentation and micro-benchmarks**: Instrumented runtime metrics (path counts, latencies, graph hit rate, queue utilization, memory pool usage) with env-configurable toggles (`CUDA_PROXY_VERBOSE/PROFILE`); built micro-bench suites and Python runners for end-to-end evaluation of launch overhead and bucketing effectiveness.
- **Edge–Cloud Collaborative VLM System for Autonomous Driving** Aug. 2025 – Dec. 2025
  CSE 589 (Advanced Networks) *University of Michigan*
  - **Network-aware AV inference pipeline**: Designed an edge–cloud collaborative VLM system combining a lightweight local VLM with a larger cloud VLM to improve out-of-distribution reasoning for autonomous driving.
  - **Network virtualization and traffic control**: Built reproducible network testbeds using Mininet, Linux network namespaces, and traffic control tools to emulate LTE-like links with controlled bandwidth, delay, and packet loss.
  - **Compression and selective offloading**: Utilize multi-level image compression and a selective offloading strategy that separates static-scene and dynamic-scene information to reduce link usage under constrained network conditions.
  - **Cloud inference server and communication stack**: Developed a cloud inference server with low-latency persistent socket communication and multi-model support, data caching.
  - **Local–cloud aggregation**: Implemented aggregation logic that fuses local and cloud reasoning signals through steering-vector fusion; evaluated accuracy using metrics such as MAE, curvature agreement, ADE, and FDE.
  - **Benchmarking and analysis**: Executed end-to-end experiments on curated autonomous-driving data with various network profiles; benchmarked compression-rate effects, offload latency, and accuracy tradeoffs.
- **Simulated Distributed System** Aug. 2025 – Dec. 2025
  EECS 491 (Introduction to Distributed System)
  - **Primary-Backup 1-Fault-Tolerant Storage System**: Implement with **Lexical Confinement** design for high concurrency request using Golang; implement the Primary Backup system that support linearizability with GET/PUT/APPEND.
  - **Paxos-Based Consensus System**: Design a 3 layer paxos replicate state machine distributed system; optimize paxos protocol safely with Accept phase skipping.
  - **Shard Distributed Systems**: Use paxos based consensus system for a replicated, high fault tolerant global view server; use paxos protocol to implement each group of storage servers.
- **Simulated Basic Operating System** Jan. 2025 – Apr. 2025
  EECS 482 (Introduction to Operating System) Lecture Project
  - **Thread Concurrency Library**: Built a lightweight user-level multicore threading library (swapcontext/makecontext): lifecycle, Mesa Monitors sync (mutex/condvar/spin), interrupts/core-suspend, non-preemptive FIFO run queues.

- **Pager & MMU**: Minimal pager (SWAP/FILE-backed); manages page tables and dirty/reference/recident bits; page-fault path: clock queue eviction, copy-on-write, defer-and-avoid; supports `fork`/`mmap`/`yield`.
- **Network File System**: Built an inode-based, Unix-style NFS with strong consistency under concurrent access; synchronized ops (create/read/write/delete) using Boost shared/unique locks; added robust error handling.
- **Linux Kernel Tracing ptrace Optimization**: Modified Linux 5.10.224 kernel to add selective memory snapshot, restore, and query support in `ptrace`.

• **Network Simulation** Jan. 2025 – Apr. 2025

EECS 489: Introduction to Computer Networks

- **Mininet simulation**: Simulated network topologies in Mininet and measured RTT/throughput with C++ sockets, also reproduce buffer bloat failure in networking.
- **Video proxy**: Built a video proxy with load balancing and adaptive DASH streaming.
- **SDN controller**: Implemented a POX SDN controller to mitigate bufferbloat by assigning traffic to QoS queues for latency-sensitive flows.
- **Transport & routing**: Implemented TCP-like reliability over UDP and an L3 router with ARP and ICMP.

• **Digital Forensics** Jan. 2025 – Apr. 2025

EECS 388 (Introduction to Computer Security) Lecture Project

- **Cryptanalysis & Cracking**: length-extension, padding-oracle; John the Ripper (PDF/ODT), Hydra (SSH).
- **Web Exploitation**: auth bypass via XSS/SQLi/CSRF.
- **Binary Exploitation**: ROP/NOP-sled against DEP/ASLR.
- **Reverse Engineering**: Ghidra decompilation and PWNing.
- **Steganography**: hidden-data detection (binwalk, Stegseek, exif check).
- **Protocols**: TLS 1.3 handshake; Google-style TOTP.

• **Auto Sentry Robot Control** 2023 – 2024

Chinese Univ. National Robot Competition – Robomaster Championship

- **Autonomous control**: Autonomous decision making and engagement with dual gimbals and 4-wheel chassis on STM32-F407.
- **Circuit & control**: Lead circuit design; dual-gimbal control stabilization; high-speed 4-wheel chassis response.
- **Perception & pose**: Developed CAN/UART pipelines for CV and LiDAR data; implemented IMU-based absolute-pose control.

Skills

- **Programming**: C/C++, Java, Rust, Golang, Python, Bash; Git; CMake, Makefile, Maven, uv, cargo
- **Systems**: Arch/Ubuntu Linux; concurrency (boost locks); MMU/paging; POSIX sockets (select/poll)
- **Networking**: tc(config); HAProxy; Mininet, POX; TCP (GBN/SR), L3 routing
- **Distributed**: Docker (Compose), Kubernetes; ZooKeeper, HDFS; Prometheus+Grafana (JMX Exporter)
- **Security**: Wireshark, Ghidra, John the Ripper, Hydra, sqlmap, Autopsy, Stegseek; ROP chains
- **ML**: PyTorch, CoT, MCP, Qwen-VLM Tuning
- **Databases**: SQLite, Oracle(SQL*Plus), MongoDB
- **Robotics**: STM32, FreeRTOS; CAN/I^2C; Flexiv RDK; PID/dynamics; C++/Rust firmware, MATLAB
- **Other**: JavaScript, HTML, Markdown, LaTeX; Neovim (LSP via Mason), SSH, tmux, GDB/LLDB

Honors & Awards

- **ICRA Best Poster Award (May 2025)**: Presented by IEEE Robotics and Automation Society.
- **Cheng–Family Scholarship (Jun. 2024)**: (Top 2%).
- **RoboMaster University Championship (May 2024)**: (Eastern Region Champion).
- **RoboMaster University League (Apr. 2024)**: (Shanghai Division Champion).
- **University Physics Competition (Nov. 2023)**: (Silver Prize).
- **SJTU Excellence Scholarship, Level B (Dec. 2023)**: (Top 3%).
- **John Wu & Jane Sun Sunshine Scholarship (Nov. 2023)**: (Top 6%).
- **RoboMaster University Championship (Aug. 2023)**: (National Champion).
- **SJTU Social Practice, Third Prize (Aug. 2023)**:

Teaching Experience

- **Teaching Assistant (2024)**: **Teaching Assistant** at Shanghai Jiao Tong University, ENGR 1000J (Introduction to Software Engineering).

Extra Curriculars

- **Undergraduate Research Assistant (May 2025 – Ongoing)**: Undergraduate Research Assistant at University of Michigan College of Engineering, MI, USA.
- **ICRA Volunteer (2025)**: **Volunteer** at IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, May 2025.
- **UM-SJTU Joint Institute Youth Volunteer Team (2023)**: UM-SJTU Joint Institute Youth Volunteer Team member (Shanghai, China).
- **Old Friends Youth Team (2023)**: Old Friends Youth Team, Shanghai, Facilitated intergenerational communication activities.

Personal Details

- **Language**: English (TOEFL 107/120), Chinese (Native).
- **Hobbies**: Badminton, Rubik's Cube, Linux Customization (Arch Linux + Hyprland + NeoVim + Fcitx5).