Shaoming Wu

647-327-1398 | s474wu@uwaterloo.ca | shaoming.vercel.app | LinkedIn | Github

Education

University of Waterloo — Waterloo, ON
BASc., Mechatronics Engineering — *Expected Graduation: May 2031*

Technical Skills

Languages: TypeScript, Python, C/C++, PostgreSQL
Frameworks/Libraries: React, Next.js, Express.js, Node.js, Django, Flutter, LangSmith, Supabase, FastAPI, Three.js

Experience

Magic Hour (YC W24) — Sep 2026 – Present
Software Engineering Intern — *San Francisco, CA*

- Accelerated **Stable-Layers** RGBA decomposition by ∼**5.7x** by stacking nine validated levers (LoRA merge, `torch.compile`, fused QKV, real RoPE, TaylorSeer, row-wise **fp8**,) and cutting **NFE 99 to 11**
- Reproduced and benchmarked **6 published diffusion-acceleration methods** (Lightning step distillation, TaylorSeer derivative extrapolation, first-block cache, fp8 quantization, token freezing, conditioning KV cache)
- Engineered a gated benchmark harness — edge-error budget anchored to a measured **0.0074** kernel noise floor, 2-seed worst-case gates, **93 tests** — exposing an accounting defect that had falsely rejected **6** approximate levers

Roboflow (YC S20) — Jul 2026 – Sep 2026
*Open-Source Contributor - Supervision (**47k+** stars)* — *San Francisco, CA*

- Vectorized hot-path label rendering and mask compositing by replacing Python loops with NumPy operations, improving annotation throughput across **5+ annotator classes**
- Optimized track cleanup in `DetectionsSmoother` by replacing O(n) linear scans with O(1) set lookups, reducing per-frame overhead for real-time **multi-object tracking** pipelines
- Authored **4 merged/open pull requests** adding vectorized routines, performance fixes, and improved test coverage, with all contributions passing **3,200+ existing tests** and **21 pre-commit hooks**

Software Engineer — Aug 2026 – Present
Hack the 6ix — *Toronto, ON*

- Architected a production **AI Slack assistant** serving **40+ organizers**, implementing a multi-round agentic tool loop over the **Anthropic API** with **7 client-side tools** and **3 server-side MCP connectors** (Notion, Outreach, Fathom) to unify org knowledge across Slack, Google Drive, and meeting transcripts
- Engineered **OAuth 2.0 + PKCE** token management for Google Drive and Notion MCP with mutex-guarded refresh, atomic writes, and rotating token persistence for zero-downtime credential rotation
- Deployed to a **free-tier Oracle Cloud** VM using a multi-stage **Docker** build, rootless **Podman** containers with systemd lingering, and a **GitHub Actions CI/CD** pipeline that auto-deploys on every push to `main` via SSH

Projects

The Brackey Way (**1st @ WAT.ai x UTMIST**) | *Python, Next.js, React, TypeScript, three.js* — Sep 2026

- Fine-tuned a π_0 **Vision-Language-Action (VLA)** model on the teleop dataset, achieving consistent bread grasp-and-place on real hardware that generalized to toaster and plate configurations unseen during training
- Ran **VLA inference** on a **remote cloud GPU**, streaming camera observations to the model and returning action chunks to the robot, with a **state machine** sequencing VLA and **OpenCV** for sandwich assembly

Hand-Sim | *Python, MuJoCo MJX, JAX, Brax, Warp* — Oct 2026

- Tuned a **MuJoCo MJX/Warp** Shadow Hand simulation to **88,106 control-steps/s** — **4,405x** realtime at **8,192** parallel environments — by capping fingertip collision hulls at **16 vertices**, **2.72x** the same config's throughput on an H200
- Diagnosed a reward-specification failure by scoring saved trajectories against the live reward: the hand-authored reference gait ranked **last of 5** behind a flick exploit by **2.3x**, traced to additive terms and a clip paying a bursty gait **50.1%** of its rotation

Awards

Hackathons (7x Winner): 1st @ JAMHacks 10, 1st @ Cursor July 2026
Other: VEX Worlds Division Top 16 (160/20,000+), Harvard Investment Pitch Finalist (10/200+)