

ALEX QIAN

Toronto, ON ☎ +1 778-636-2337

✉ lihao.qian@outlook.com

🐙 GitHub

🌐 LinkedIn

📊 Portfolio

Education

University of Waterloo

M.Eng., Electrical Computer Engineering

Sep 2025 – Aug 2026

Waterloo, ON

Simon Fraser University

B.Sc., Computer Science

Jan 2020 – Dec 2024

Burnaby, BC

Technical Skills

- **Languages:** C/C++, Python, Assembly, Bash, TypeScript/JavaScript, Java, Kotlin, Swift, Matlab, SQL
- **AI & ML:** PyTorch, Hugging Face, scikit-learn, NumPy/CuPy, pandas, OpenCV DNN, YOLOv8/v11, CNNs & Transformers
- **Inference Efficiency:** 4-bit quantization (bitsandbytes), vLLM serving, ONNX/TFLite/CoreML export, real-time C++ inference
- **GPU & HPC:** CUDA, cuVS/CAGRA, CuPy, multi-GPU sharding, zero-copy shared memory, multithreading, MPI, SLURM
- **Embedded:** BeagleBone Green (Embedded Linux), Arduino, Raspberry Pi, PRU, I2C/GPIO/UDP, sensor & display drivers, Qt
- **LLM & MLOps:** LLM agents, RAG, hybrid retrieval (BM25 + FAISS), eval harnesses (Recall@K, MRR, EM/F1), Docker, CI/CD, pytest, FastAPI, AWS

Professional Experience

Puncture Robotic

May 2025 – Aug 2025

Software Engineer Intern

Shanghai, China

- Built real-time follicle detection-and-tracking module in C++/Qt for a **robotic hair-transplant surgery** system: **YOLOv11 OBB + instance segmentation** via **OpenCV DNN**, **SIFT/ORB** matching fused with **Lucas-Kanade/RLOF** optical flow, and a from-scratch **DBSCAN** to cluster follicles into single/multi-unit groups for distribution analytics.
- Invented a multi-factor follicle-root correction algorithm in C++/**OpenCV**, named **sole inventor** on the resulting **corporate patent filing**, fusing sliding-window gradient analysis, Gaussian intensity scoring, and distance penalties, lifting localization accuracy **+42.7%** over the single-threshold baseline to sub-pixel **RMSE <0.8 px** at **<3 ms/point**.
- Built a stereo-vision **3D reconstruction** pipeline: epipolar rectification, disparity maps, and reprojection into RGB-textured **point clouds**, refined via **Voxel-grid** downsampling, **FPFH+GICP** registration, and **Poisson/NURBS** meshing; benchmarked **CUDA SGBM** on GPU and identified the 256-disparity **VRAM** ceiling in libSGM, shipping an ROI-cropped CPU path.

Simon Fraser University

May 2024 – Aug 2024

Research Assistant, ixLab

Burnaby, BC

- Researched **interactive natural materials** in SFU's **ixLab**, characterizing novel chemical and physical processes to render solid wood panels optically **transparent** and electrically **conductive**.
- Prototyped a **Raspberry Pi**-driven color display embedded beneath the treated wood surface and **capacitive touch** sensing through the solid panel via an **MPR121** controller, turning an ordinary wood panel into a seamless **touch-interactive display**.

Project Experience

Review++: Local-First AI Study Assistant (Electron, React, TypeScript, LLM Agents, Context Engineering, OAuth 2.0)

- Built an **Electron/React TypeScript** assistant on a **tool-calling agent harness**: **11** typed tools in bounded multi-round loop, runtime **function-calling** detection with retrieval-only fallback, across **6** LLM providers spanning **Anthropic&OpenAI** formats.
- Replaced whole-corpus prompt injection, capped at **80K** characters and dropping files past it, with a budgeted **context assembler** over hybrid **BM25**/embedding retrieval, cutting per-turn context **16x**; history **compacts incrementally** into a running summary.
- Engineered a knowledge base over **17** file formats: from-scratch **OOXML** parsing, **Tesseract OCR** fallback, content-hash caching; enforced an **agent security boundary** where model-supplied tool arguments resolve only against indexed keys.

Gatekeeper Insights (Embedded C, BeagleBone, Arduino, Hardware-Software Integration, IoT Protocols)

- Developed an embedded **customer traffic monitoring system** in **C** to track real-time store occupancy and traffic trends, streaming live counts over **UDP** to a **Node.js** dashboard with **IndexedDB** storage.
- Integrated **BeagleBone Green (Debian Linux)** with **Arduino**, writing **C drivers** for motion sensors, laser-photocell gates, **MCP9808** temperature sensors, and **PRU**-driven **NeoPixel** LEDs.
- Designed **joystick-controlled mode switching** between **People Count**, **Temperature**, and **Smile** modes on an **8x8 LED matrix**, with an **I2C GPIO** 14-segment occupancy display and a 4-digit Grove clock.

LiveLens: Live-Event Seat-Review Platform (React, TypeScript, FastAPI, PostgreSQL, AWS, LLM, Three.js)

- Built a full-stack **live-event seat-review platform** where users search venues, events, and seats and post multi-dimensional view, sound, and value **ratings**, backed by a **FastAPI + PostgreSQL/PostGIS** API with filtered geospatial search, seat-level rating aggregates, and **JWT** plus **Google OAuth** authentication.
- Shipped **AI** features including a **GLM-4/Gemini** function-calling assistant that answers natural-language questions through **10** database-query tools, plus **text-to-image** seat-view generation and automatic review-tag extraction.
- Delivered a **React/TypeScript** frontend with an interactive 3D venue viewer, deployed on **AWS App Runner**, **RDS**, and **Amplify/CloudFront**, provisioned via **AWS CDK** with **CI/CD** auto-deploys and **Selenium/pytest** test coverage.

Multi-GPU Approximate Nearest Neighbor Search (Python, CUDA, NVIDIA cuVS/CAGRA, CuPy, Multiprocessing, HPC)

- Extended the single-GPU **CAGRA** graph ANN algorithm into **multi-GPU** vector-search framework on a sharded scatter-gather architecture, using **multiprocessing** to bypass the **GIL** for true parallel index construction across GPUs on a **SLURM** cluster.
- Built a **zero-copy** loading pipeline over **shared memory** and **CuPy** that streams **SIFT1B** vectors into a single shared buffer, indexing **50M+** vectors without host-RAM duplication and **2.5×** the single-GPU capacity.
- Cut 20M-vector index build time from **1884s** to **211s**, an **8.9×** **speedup**, while holding **>97% Recall@10** at scale through dynamic global top-k merging across partitions.