Vaibhav Attre

Systems & ML Infra | C/C++/Python | Kernel Optimization, Inference, Full-Stack
vattre@uci.edu | +1 (425) 365 - 9557 | GitHub | LinkedIn

Summary
Junior at UC Irvine specializing in **systems, ML inference, and performance engineering**. Shipped production-grade work across **AI accelerator kernel optimization** (Tenstorrent Blackhole), **edge–cloud inference pipelines**, and **full-stack ML systems**. Comfortable from bare-metal OS internals to cloud infra and end-to-end product delivery.

Education

University of California, Irvine — Irvine, CA
B.S. Computer Science | Dean's Honor List | GPA: 3.707 — Expected March 2027
Relevant Coursework: Operating Systems, Computer Architecture, Data Structures & Algorithms, System Design, ML

Technical Skills

Languages: C, C++, C#, Java, Python, TypeScript, Verilog
AI / Inference: PyTorch, TorchScript, YOLO, Gemini API, latency analysis, memory layout & data movement
Systems/Tools: Linux, Git, Docker, QEMU, log/trace analysis
Full-Stack: FastAPI, Next.js, PostgreSQL, REST APIs, Supabase
Cloud/Infra: AWS (EC2, S3, SQS, RDS, Cognito, Amplify, DynamoDB, IAM), GitHub Actions CI

Experience

TuriyamAI | Software Development Intern — June 2025 – September 2025 | Redmond, WA
- Improved throughput by **12%** on selected kernels for **Tenstorrent Blackhole** by optimizing **NoC-aware data movement** and specialized functional units; profiled **8** custom PyTorch operators and identified critical bottlenecks.
- Reduced kernel latency by **9%** via **tiling** and **double buffering**; cut stall time by **20%** in targeted execution traces.
- Built repeatable **perf + validation harnesses**: standardized configs, automated **60-run** batches, regression checks, and a results summary adopted across the team.

GoFlyy | ML Systems Intern (Computer Vision Pipeline) — January 2026 – Present | Irvine, CA
- Designed and shipped an end-to-end **clothing scan ML pipeline**: built a scoring system using **Gemini** for initial classification, currently migrating to **YOLO** with a custom-model roadmap; validated effectiveness across **9 garment categories** via **100+ targeted regression tests**.
- Architected the full backend: **FastAPI** + **Next.js** frontend, **PostgreSQL** on AWS RDS (Docker for local dev), **S3** artifact storage, async job queue with traceable DB/API flows to support auditability and debugging.
- Integrated **AWS Cognito, Amplify, and RDS** for auth, deployment, and managed database; designed schema and API contracts to decouple ML scoring from ingestion for independent iteration.

UCI Undergrad Research | Researcher (ML Systems) — January 2026 – Present | Irvine, CA
- Developing **runtime guardrails** for early-exit / split inference to maintain QoS (accuracy + latency SLA) under varying wireless conditions; instrumented pipeline logging bailout events over **1000+** runs to inform adaptive thresholding strategy.

Idori | Software Development Intern — January 2025 – June 2025 | Irvine, CA
- Increased average FPS by **35%** in constrained WebGL via **GPU instancing**, batching, and shader simplification (reduced frame-time variance by **18%**); eliminated race/crash incidents from **5+/week** to **<1/week** via thread-safe producer–consumer pipeline (mutexes/semaphores).

Projects

TinyOS — RISC-V OS + Disk-backed CoW Filesystem | C, RISC-V, QEMU — 2025 – Present
- Built an RV64 OS from scratch: **Sv39 page tables**, page-fault handling, trap/syscall entry/return, timer-interrupt preemption, and round-robin scheduling; passes **45+** user-level tests with full `exec`/`init` support.
- Implemented a disk-backed **copy-on-write filesystem**: VirtIO block driver + buffer cache, checksummed B-tree metadata, atomic commits, extent allocation, refcounted CoW writes, snapshots/subvolumes + reflink `clone`.
- Hardened reliability via invariant checks and **100+** iteration alloc/CoW stress loops to surface subtle memory bugs.

SentinelQ — Edge–Cloud Home Surveillance | Python, C++, Next.js, FastAPI, PostgreSQL — March 2026
- Built a low-cost surveillance platform on **Qualcomm Arduino UNO Q** handling **4+ simultaneous camera streams**; on-device model achieved **84% accuracy**, routing to cloud for deeper analysis at **90% accuracy**.
- Designed an **adaptive inference pipeline** with a custom **branchy ResNet**: exits locally or escalates to cloud based on connectivity, camera quality, and latency constraints – minimizing bandwidth while preserving detection quality.
- Shipped full-stack product: **Next.js** frontend + **FastAPI** backend + **PostgreSQL** (Supabase), on-device HTTP debug endpoint (live feed + metrics), and **LLM control assistant** for text-command threat summaries.

Telemetry Platform — Cloud QEMU Experiment Runner | Python, QEMU, Docker, AWS — 2026 - Present
- Automated QEMU workload runs to collect artifacts (`kernel.log`, `run_meta.json`, `metrics.json`) and diff runs to detect regressions; processed **200+** runs across **10** workloads.
- Built an AWS-backed queued runner: SQS jobs → EC2 worker service runs Dockerized QEMU → S3 artifacts + DynamoDB indexed metadata for querying and traceability.
- Added **GitHub Actions** CI to run smoke/regression workloads on pushes and publish run outputs.

Activities & Awards

WebJam Hackathon (3rd Place) – Built a real-time AI PhotoBooth web app (computer vision + live filters) **AI Club** – Course planning assistant
Video Game Design Club – Enemy AI behaviors and modular ability system **AWS Club @ UCI** – Event Operations Lead