# Dilraj Devgun

New York, NY · dilrajsinghdevgun@gmail.com · 425-279-3444 · www.thejarlid.com

## Education

**University of Washington** — Seattle, WA
BS Computer Science *GPA: 3.75* — Aug 2015 - Jun 2019

## Experience

**Meta** — New York, NY
*Senior Software Engineer - XR Eyes* — Oct 2023 - Present

- Productionized advanced Eye Tracking CV/ML solutions for next-generation AR/VR devices, serving as the primary owner of the C++ Eye Tracking engine to enable real-time, on-device inference
- Profiled CPU, memory, and power usage of Eye Tracking to find bottlenecks and optimize image processing through both vectorized SIMD implementations of OpenCV algorithms and hardware acceleration, achieving sub-10ms end-to-end latency for real-time gaze tracking
- Partnered with ML Researchers to enable novel model architectures and capabilities in production by bringing up Executorch (PyTorch Edge) runtime support and complex temporal frame sequencing
- Designed public OS-level APIs enabling first and third-party developers to access real-time gaze data streams for system input, display optimizations, and interactive applications
- Drove end-to-end feature development and collaboration across firmware, product, and ML teams to pioneer Eye Tracking integration as a real-time signal for Meta's contextual AI/LLM stack
- Debugged ML model generalization issues during deployment by investigating tensor parity between on/offline execution, identifying training data problems and resolving them through image augmentation
- Redesigned our inference engine into a modular, product-agnostic architecture with modern software patterns, clean abstractions, and interfaces to accelerate bring-up across multiple AR-like devices

*Senior Software Engineer - Instagram Web Server* — Oct 2022 - Oct 2023

- Enhanced the performance, scalability, and reliability of Instagram's infrastructure to support billions of global users as part of Web Server team
- Improved IG Django server fleet efficiency by >3% by implementing immortal instances into CPython runtime and optimizing refcount behavior; this is now upstream in official CPython source (PEP-683)
- Optimized threading and fiber memory on IG's server fleet by 66% (600MB/Host) by profiling system memory allocations and fine-tuning threadpool sizes
- Worked with production engineering teams to improve load balancing mechanisms during high traffic scenarios reducing timeouts and machine pressure

*Senior Software Engineer - AR Glasses CV/ML Firmware Team* — Apr 2021 - Oct 2022

- Developed custom silicon firmware to accelerate on-device Computer Vision and Machine Learning for Orion and future AR glasses
- Adapted Structured Light Depth algorithms from research implementations into high-performance production C/C++ to produce depth maps from raw sensor data in an embedded system

**Microsoft** — Redmond, WA
*Software Engineer* — Aug 2019 - Apr 2021

- Software development for custom ASIC on the HoloLens 2 and IVAS Project
- Tripled frame rate for IVAS cameras by implementing shared memory buffer libraries in an embedded memory constrained environment
- Maintained C++ tools for recording and replaying sensor data to debug and evaluate runtime algorithms

## Projects

**Vibe Dine** — www.vibedine.ai

- Built a full-stack application enabling users to query their saved google maps places through natural language using a Next.js/React frontend and a FastAPI backend
- Implemented an ingestion pipeline to scrape and index data into Vector and PostgreSQL Databases for agentic search/retrieval, with LangGraph orchestration, GPT/Claude, and custom tools for web search

## Skills

Proficient Programming Languages: C, C++, Java, Python, Swift, Objective-C, C#