

Brian Minh Phan

brian.phan@columbia.edu | (408)-757-8291 | www.banebeetle.com | www.linkedin.com/in/banebeetle

EDUCATION

Columbia University

Master of Science: Computer Science

Expected Dec 2026

New York, NY

- Coursework: Machine Learning, Deep Learning, Natural Language Processing, Reinforcement Learning

University of California, Irvine

Bachelor of Science: Computer Science

Jun 2025

Irvine, CA

EXPERIENCE

Machine Learning Engineer

Calit2

Oct 2024 — Mar 2025

Irvine, CA

- Built real-time LSTM-based prediction model in PyTorch simulating large-scale thermal behavior of energy storage systems; reduced simulation time from 15 hours to 6 minutes (including training)
- Engineered a custom training pipeline with gradient clipping and teacher forcing, effectively mitigating exploding gradients and stabilizing convergence across 100 epochs of volatile time-series data

Guest Lecturer & Teaching Assistant

University of California, Irvine

Sep 2023 — Mar 2025

Irvine, CA

- Guest-lectured a 360-student session on Python and Data Structures; led a team of learning assistants using Agile workflows to coordinate grading, exams, and administrative operations
- Developed a full-stack LLM-powered quiz generator using Python, FastAPI, and React on AWS EC2 with rate limiting and input sanitization, reducing content creation workload by 93%

PROJECTS

Iron Bark (Autonomous VLA Robot Dog) | Multi-Modal Perception, Edge AI

Mar 2026 — Present

- Architected a multithreaded perception pipeline combining YOLOv11 object detection, multi-shot ArcFace biometric enrollment, and a Vision-Language Model (VLM) a distributed Pi 5 + GPU compute architecture
- Eliminated VLM hallucinations by designing YOLO-grounded prompting, injecting real-time detection counts into VLM context, and swapping LLaVA-7B for Moondream, cutting latency 4.7x
- Developed a behavior state machine (IDLE → FOLLOW → SEARCH → EXPLORE) with hysteresis-based transitions driven by VLM reasoning over a dual-camera distributed edge system using ZeroMQ

Rembeetle.com (Real-Time Agentic Voice Platform) | Generative Models, AWS EC2

Jul 2025 — Present

- Designed a real-time multimodal inference pipeline with sub-second latency using asynchronous Python, orchestrating Whisper ASR, GPT-4o, TTS, and Silero VAD for bidirectional voice AI
- Built a Retrieval-Augmented Generation (RAG) system with Google OAuth (ES256/JWKS), asyncpg Postgres with per-user data isolation, and pgvector cosine similarity search over 1536-dim embeddings for persistent semantic memory
- Engineered an Agentic AI pipeline using the Model Context Protocol (MCP) with streaming JSON parsing, enabling autonomous tool discovery and real-time function calling across external services

Autonomous Voice-Controlled Robotic Arm | YOLOv11, OpenCV, UDP Networking

Sep 2025 — Nov 2025

- Engineered a 4-DOF autonomous robotic arm using a custom-trained YOLOv11 object detection model to execute real-time voice commands via computer vision
- Designed a deterministic control state machine with 7-point linear interpolation to map camera depth to precise servo angles with 0.5-degree actuation accuracy
- Built a distributed control architecture processing vision at 8 FPS on a host PC and streaming coordinates via low-latency UDP to a Raspberry Pi embedded controller

TECHNICAL SKILLS

ML & AI: PyTorch, TensorFlow/Keras, Scikit-learn, OpenCV, YOLO, Transformers (Hugging Face), LSTM/RNNs, Deep Learning, Computer Vision, Object Detection, Neural Networks, Transfer Learning, Inference Optimization, RAG, CUDA

Programming Languages: Python, C++, C, SQL

DevOps & Cloud: AWS (EC2, S3), Git, Linux, REST APIs, WebSockets, OAuth/JWT, Supabase

Robotics & Hardware: Raspberry Pi, Embedded Systems, Actuator Control, State Machines, Real-Time Systems, Distributed Systems, Multithreading, ZeroMQ, UDP Networking

Data & Databases: MySQL, PostgreSQL, asyncpg, Pandas, NumPy, pgvector