

## HENRY NGUYEN

San Jose, CA | (206) 751-6618 | henrynguyen.vp@gmail.com

Website: henrynvp.github.io | GitHub: github.com/HenryNVP | LinkedIn: linkedin.com/in/henrynguyen-vp

### EDUCATION

#### **Master of Science in Artificial Intelligence**

San Jose State University

Exp. May 2026

San Jose, CA

- Coursework: AI & Data Engineering, Deep Learning, Reinforcement Learning, Autonomous Driving, MLOps

#### **Bachelor of Engineering in Mechatronic Engineering**

Ho Chi Minh City University of Technology

Apr. 2023

Vietnam

### TECHNICAL SKILLS

**AI & Agents:** LLMs, LangGraph, GraphRAG, MCP, Vector/Graph DBs, Agentic Workflows

**Machine Learning:** PyTorch, TensorRT, ONNX, CUDA, Quantization, Transformers, CNNs, RecSys

**Robotics:** ROS2, LiDAR-Camera Fusion, 3D Perception, Navigation, Control Systems, System Integration

**Software & DevOps:** Python, FastAPI, Docker, Microservices, Prometheus, Edge AI (Jetson, ONNX Runtime)

### FEATURED AI PROJECTS

#### **AI Tutor: RAG-Powered Learning Platform** | *GenAI, RAG, MCP, FastAPI, OpenAI Agents SDK*

- Built a full-stack **multi-agent** educational system that ingests documents to generate cited answers, adaptive quizzes, and lesson notes via a source-filtered **RAG** pipeline using **ChromaDB**.
- Implemented an **MCP server** and secure Python execution with **FastAPI** backend, enabling structured tool use, real-time data visualization from CSVs, and adaptive learning features that track student progress.

#### **ROS2 BEV-Fusion: Real-Time 3D Perception** | *Python, ROS2, TensorRT, CUDA, Jetson, Edge AI*

- Developed an optimized **BEVFusion** 3D perception pipeline for multi-camera and LiDAR fusion, validated on NuScenes and deployed as a modular **ROS2** package.
- Optimized end-to-end inference with **TensorRT** and quantization, achieving ~7 FPS for the full BEVFusion pipeline on Jetson Orin Nano and publishing **ROS2** detection outputs with latency metrics.

#### **FastViT Mobile Optimization** | *PyTorch, Android, Quantization, Knowledge Distillation*

- Re-architected FastViT by replacing Multi-Head Attention with **Performer Attention** ( $O(N)$ ) and implementing **FP16 quantization**, achieving a **4.8x speedup** on Android with **identical Top-1 accuracy**.

#### **SAM-E: Agentic Enrollment System** | *GenAI, RAG, LangGraph, Docker, FastAPI*

- Architected a microservices-based agentic system with three services (Agent, RAG, Enrollment Engine) using **Docker Compose** and **LangGraph** to route user intents to specialized tools.
- Developed a retrieval pipeline using **pgvector** to support academic queries, with planned integration of a **Neo4j** knowledge graph; demonstrated functionality via **FastAPI**, **JWT authentication**, and **Prometheus** metrics.

### ADDITIONAL PROJECTS

**Image Classification:** Engineered a modular **timm** training pipeline with automated **ONNX** export, enabling rapid benchmarking of CNN/ViT architectures.

**3D Object Detection Pipeline:** Built **MMDetection3D** end-to-end inference pipeline for KITTI/nuScenes.

**Anime RecSys:** Trained and deployed **NeuMF** and **Two-Tower** recommender system via **FastAPI**.

**Client Web Projects:** Delivered commercial **WordPress** sites with automated booking, increasing client inquiries.

### PROFESSIONAL EXPERIENCE

#### **Software Engineer (Automotive Systems)**

Bosch Global Software Technologies

Jun. 2023 – Dec. 2023

Ho Chi Minh, Vietnam

- Delivered production-ready **ESP** software for 8+ projects and provided technical support for several others, leading **ISO-compliant** integration testing and accelerating defect resolution via optimized **HIL** workflows.

#### **Undergraduate Researcher**

Mechatronics Lab, HCMUT

Aug. 2022 – Dec. 2022

Ho Chi Minh, Vietnam

- Designed an adaptive 3-finger robotic gripper (+200% payload capacity) and developed the C++ control stack for a 5-axis manipulator to execute automated pick-and-place tasks.