HENRY NGUYEN

San Jose, CA | (206) 751-6618 | henrynguyen.vp@gmail.com
Website: henrynvp.github.io | LinkedIn: linkedin.com/in/henrynguyen-vp | GitHub: github.com/HenryNVP

EDUCATION

Master of Science in Artificial Intelligence *Exp. May 2026*
San Jose State University San Jose, CA
- Coursework: AI & Data Engineering, Machine Learning, Deep Learning, Intelligent Autonomous Systems, Data Mining

Bachelor of Engineering in Mechatronic Engineering *Apr. 2023*
Ho Chi Minh City University of Technology Vietnam

TECHNICAL SKILLS

AI & Agents: LLMs, LangGraph, RAG, Vector DBs, LangChain, Agentic AI, Reasoning techniques (Chain-of-Thought).
Machine Learning & Deep Learning: PyTorch, ONNX, TensorRT, Scikit-learn, CNNs, Transformers, Recommender systems.
Robotics & Autonomous Systems: ROS2, LiDAR-Camera Fusion, SLAM, Control systems, Autonomous driving pipelines.
Software & Deployment: Python, FastAPI, Microservices, Git, Docker, Edge & mobile AI deployment.

FEATURED AI PROJECTS

SAM-E: Multi-Agent Enrollment Assistant | *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.

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 $\sim$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*
- Engineered a mobile-optimized FastViT architecture by replacing Multi-Head Attention with **Performer Attention** to reduce complexity to $\mathbf{O(N)}$, and implementing an **FP16 quantization** pipeline.
- Achieved a **4.8x inference speedup** on Android devices compared to FP32 baselines while maintaining **identical Top-1 accuracy**, validated via a custom Android benchmarking app.

ADDITIONAL PROJECTS

3D Object Detection Pipeline: Built **MMDetection3D** inference pipeline for KITTI/nuScenes, achieving $\sim$12 FPS throughput on Tesla P100.

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

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

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

PROFESSIONAL EXPERIENCE

Software Engineer (Automotive Systems) Jun. 2023 – Dec. 2023
Bosch Global Software Technologies Ho Chi Minh, Vietnam
- Led integration testing for the Electronic Stability Program (ESP) across 10+ projects, ensuring ISO-standard compliance for safety-critical automotive software.
- Collaborated with 5+ cross-functional teams to debug Hardware-in-the-Loop (HIL) failures, reducing defect resolution time by streamlining the error-reporting workflow.

Undergraduate Researcher Aug. 2022 – Dec. 2022
Mechatronics Lab, HCMUT 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.