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 *Exp. May 2026*
San Jose State University San Jose, CA

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

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

TECHNICAL SKILLS

Languages: Python, C++, C, SQL, Bash, MATLAB
AI & Agents: LLMs, LangGraph, RAG, MCP, Vector/Graph DBs, Agentic Workflows
Machine Learning: PyTorch, TensorRT, ONNX, CUDA, Quantization, Transformers, CNNs
Robotics: ROS2, System Integration, Sensor Fusion, Navigation, Control Systems
Software & Tools: Git, Docker, CI/CD, Linux, FastAPI, Microservices, Edge AI (Jetson)

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 **MCP servers** 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 | *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, ONNX Runtime, Quantization, Knowledge Distillation*

- Re-architected FastViT by replacing Multi-Head Attention with **Performer Attention** ($\mathbf{O(N)}$) 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 **timm** pipeline with automated **ONNX** export for rapid CNN/ViT benchmarking.

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 Integration Engineer (Automotive Safety Systems) Jun. 2023 – Dec. 2023
Bosch Global Software Technologies Ho Chi Minh, Vietnam

- Managed the software release lifecycle for **8+ ESP (Electronic Stability Program)** projects, integrating modules from cross-functional teams to deliver production-ready baselines.
- Executed comprehensive integration testing (**SiL & HiL**) and authored **ISO-compliant validation reports**, identifying critical defects to ensure system stability before delivery.

Undergraduate Researcher Aug. 2022 – Dec. 2022
Mechatronics Lab, HCMUT Ho Chi Minh, Vietnam

- Designed and fabricated a custom 3-finger gripper (+200% payload) and **retrofitted** a 5-axis manipulator, engineering a **vision-guided control stack** (C++, YOLO) for automated pick-and-place.