Ho Thinh Hung

hothinhhung77@gmail.com | +84 3533 02886
GitHub | LinkedIn | Blog | Portfolio

Research Profile

My research focuses on **efficient and robust embodied intelligence**: how vision-language-action policies can retain task performance under limited memory, computation, and distribution shifts. I work on compact VLA development and training, native low-bit inference, and retrieval-augmented robot execution, connecting algorithmic research with reproducible policy evaluation and deployment on edge hardware.

Research advisor: Dr. An Thai Le — Assistant Professor, VinUniversity; Director of Foundation AI, VinRobotics.
Research interests: Robot learning; vision-language-action models; model quantization; efficient inference; retrieval for robust robotic execution.

Education

HCMC University of Science, VNU-HCM Aug 2016 – Dec 2020
B.S.E. in Computer Science
Thesis: *3D Object Detection on LiDAR Point Clouds and Applications.*

- Adapted YOLOv3 for oriented 3D object detection from LiDAR bird's-eye-view maps encoding point density, height, and intensity.
- Integrated 3D SORT with Kalman filtering and Hungarian matching for tracking; evaluated detection on 1,122 KITTI samples and tracking using 3D IoU and positional error.

Publications & Manuscripts

FoldQuantVLA: Native Low-Bit Quantization of Vision-Language-Action Models via Consistent Folding
Hung T. Ho et al. | First author | Under review, 2026
arXiv:2609.24433 | Code

- W8A8 and W4A4 quantization of vision-language-action (VLA) models, executed natively on the device's INT8 / INT4 tensor cores, via consistent offline folding.

vla.cpp: A Unified Inference Runtime for Vision-Language-Action Models
Khanh D. Nguyen, **Hung T. Ho**, et al. | Co-author | Under review, 2026
arXiv:2606.08094 | Code

- Unified C++ execution of VLA policies with portable model packaging and heterogeneous inference backends, bridging research models and practical robot deployment.

Retrieve to Act: Motion Primitive Graph Retrieval for Robust VLA Execution
Co-author | Under review, 2026 | Project page

- Research retrieval-augmented robotic execution using motion primitive graphs to support robust VLA behavior under out-of-distribution (OOD) conditions.

Research & Engineering Experience

VinRobotics Jan 2026 – Present
Senior Edge AI Engineer — Robotics Research

- **VLA optimization & deployment framework:** Develop a framework spanning model optimization, export, runtime integration, and deployment-oriented evaluation of memory, latency, and policy behavior.
- **FoldQuantVLA:** Research native W4A4/W8A8 VLA inference through consistent offline folding, calibration, and weight rounding; integrate integer execution paths for language backbones and action experts using TensorRT/CUTLASS.
- **vla.cpp:** Contribute to a unified C++ runtime for VLA policies across heterogeneous hardware, supporting deployment without Python or PyTorch at inference time.
- **Model Quantization Recipes:** Build reproducible FP8, NVFP4, INT8, and INT4/AWQ workflows with NVIDIA ModelOpt, export tooling, and deployment-oriented validation for LLM and ASR systems.
- **EdgeVLA-TRT:** Develop a VLA/VLN inference framework built on NVIDIA TensorRT Edge-LLM, extending model export, engine build, and C++ policy serving on Jetson. Integrate models including GR00T, SmolVLA, X-VLA, and OpenVLA, with policy validation in LIBERO simulation on AGX Orin.
- Extend the runtime with asynchronous action chunking, dual-rate planner execution, and KV-cache reuse; add JetPack 6 support and address numerical, AWQ repacking, and model-export failures through layer-by-layer validation.

- Submitted **NVIDIA PR #193** for InternVLA-N1-DualVLN navigation support, connecting a vision-language planner with a trajectory expert and validating intermediate latent outputs on Jetson Thor.

FPT Software AI Center Oct 2024 – Jan 2026
AI Engineer *Ho Chi Minh, Vietnam*

- Built **Camelot smart parking analytics** with YOLO11 vehicle detection, BoostTrack multi-object tracking, and cached tracker/occupancy state for continuous video processing.
- Curated Vietnamese healthcare data and fine-tuned a domain LLM; evaluated generated answers using automatic metrics and expert assessment.
- Deployed **Qwen3 14B on an 8-H100 setup** using vLLM and LMCache, exposing OpenAI-compatible blocking and streaming APIs through SSE/WebSocket.
- Designed a sketch-to-3D CAD pipeline combining sketch understanding, depth/shape estimation, and geometry-aware constraints.

FPT Corporation Mar 2022 – Oct 2024
AI Engineer *Ho Chi Minh, Vietnam*

- Built GPU-accelerated face recognition for FPT Telecom, combining AdaFace embeddings, low-quality image enhancement, and image quality assessment.
- Optimized inference with **CUDA, ONNX, TensorRT, and DeepStream/C++**; supported **100+ concurrent Full HD streams per server** through parallel processing and concurrency control.
- Developed a multithreaded C/C++ detection-and-tracking pipeline for retail people counting, achieving **93% counting accuracy** in the Long Chau Pharmacy deployment.
- Prototyped contextual advertising with BLIP-2 scene descriptions, open-vocabulary detection, and moment segmentation for video understanding.

GMO – Z Runsystem Jun 2019 – Feb 2022
AI Engineer *Ho Chi Minh, Vietnam*

- Developed Vietnamese document extraction with PICK and handwritten text recognition using DB detection, CRNN, and CTC loss, including support for Vietnamese diacritics.
- Curated printed/handwritten datasets, applied augmentation, and deployed document processing through REST APIs; achieved **85% passport information extraction accuracy**.

Personal Project

Multi-Agent Digital Human Virtual Reception

- Built a WebRTC-based virtual receptionist combining streaming ASR, VAD, dialogue, neural TTS, lip-sync, and avatar rendering.
- Orchestrated perception, dialogue, task, and avatar agents using Kafka/Redis and typed event contracts; integrated hybrid BM25/dense retrieval with FAISS/pgvector for grounded responses.
- Containerized GPU inference with Docker and CUDA/TensorRT; incorporated ASR/TTS fallback paths for service continuity.

Technical Skills

Languages & ML: Python, C++, CUDA; PyTorch, TensorFlow, Hugging Face.
Robotics & Multimodal: Compact VLA model development, VLA training, vision-language models, action experts, retrieval-augmented robot execution, asynchronous policy inference.
Optimization: PTQ, calibration, mixed precision, FP8/NVFP4, INT8, INT4/AWQ, SmoothQuant, GPTQ, NVIDIA ModelOpt; numerical validation and memory/latency profiling.
Inference & Hardware: TensorRT, TensorRT Edge-LLM, ONNX Runtime, vLLM, LMCache, Triton Inference Server; NVIDIA Jetson and GPU systems.
Perception & Systems: DeepStream, YOLO, tracking, face recognition, OCR; Docker, Kafka, Redis, MongoDB, WebRTC, REST, WebSocket.

Honors & Awards

QAI Hackathon finalist (2022); Employee of the Quarter (Q4 2022, Q3 2023); Hall of Fame recognition (July 2025).