

THAN VAN HONG SON

AI Engineer | GenAI / LLM | Computer Vision | MLOps

Da Nang, Vietnam | 0974 830 945 | sontvh2002@gmail.com | linkedin.com/in/than-hong-son

AI Engineer with **3+ years shipping production AI**: multi-agent LLM pipelines running fully on local GPUs, real-time computer vision, and end-to-end MLOps. Currently building a **real-time multilingual meeting assistant at FPT Software** while leading a part-time AI team at BK High Tech. First author of a **Springer publication**. Strongest at turning research-grade models into products that run **fast and cheap in production**.

WORK EXPERIENCE

AI Engineer | FPT Software

Jun 2026 - Present | Da Nang | Full-time | Real-time meeting-intelligence platform

- Build the AI core of a **real-time meeting assistant**: live transcription, **speaker diarization** and meeting Q&A.
- Designed the **streaming-quality benchmark** for STT + translation (CER, BLEU, COMET/XCOMET, **StreamLAAL**, RTF) on **SimulEval**.

AI Lead (Part-time) | BK High Tech

Jul 2025 - Present (part-time since Jun 2026) | Remote | Comic AI Translation

- Lead a cross-functional AI team (task allocation, resources, roadmap) building an **end-to-end comic localization pipeline**: **OCR → translation → inpainting**, replacing most manual processing for localization teams.
- Designed a **multi-agent orchestration system (LangChain)** that improves translation quality and consistency using **fully local LLMs**, with no external API dependency.
- Brought cloud inference cost to **zero**: quantization + VRAM management to run the entire stack on **constrained consumer GPUs**.

AI / Product R&D Specialist | Permate

Apr 2024 - Jan 2026 | Da Nang | Performance-marketing platform (permate.com)

- Designed the company's **MLOps platform** (Airflow, MLflow, Feast, Grafana, Jupyter, Docker), standardizing the train → deploy → monitor loop for all ML workloads.
- Researched and deployed a **Lead Management System for a US-market FinTech** R&D product.
- Built **fraud-prevention and customer-journey analytics** on browser time-series data, incl. **cross-visit device identification**; owned client integrations & docs (DevRel).

AI Engineer - Generative AI | BK High Tech (Pixor AI)

Sep 2023 - Mar 2024 | Da Nang | Stable Diffusion text-to-image product (pixor.ai)

- Shipped ML features for a consumer image-generation app: **hashtag classification (MobileNetV2)**, **face recognition (ResNet50)** and **face swapping (SimSwap)**.

AI Engineer Intern - Computer Vision | Smartinside AI

Mar 2023 - Aug 2023 | Remote, Korea | AI monitoring for live construction sites

- Improved **multi-camera MOT tracking & ReID** with **NVIDIA DeepStream** for real-time site safety: fall detection, worker-trajectory prediction, crack-length estimation, point-cloud mapping.

CORE SKILLS

GenAI / LLM

LangChain, multi-agent systems, local LLM deployment, quantization, prompt engineering, Stable Diffusion

Computer Vision

PyTorch, TensorFlow/Keras, MOT & ReID, DeepStream, OCR, pose estimation, face recognition

MLOps & Data

Airflow, MLflow, Feast, Docker, Grafana, SQL/NoSQL, time-series analysis

Languages

Python, C/C++, C#, JavaScript, R

EDUCATION

M.Sc. Computer Science

Da Nang University of Science & Technology (DUT)
Oct 2025 - Present

B.Sc. Data Science & AI

Da Nang University of Science & Technology (DUT)
2020 - 2024 | GPA 3.47/4.0 (Very Good)

PUBLICATIONS

FIRST AUTHOR | SPRINGER

Real-Time Detection and Evaluation of Oranges on Mobile Devices, Proc. 1st Int'l Conf. SETSM 2024.

CO-AUTHOR

Gym-Gesture Recognition Combining Transformer and 3D Pose Estimation, Cybernetics and Physics Journal, 2024.

HONORS & AWARDS

Top 2 - TAS Design Competition (NLP) | Dec 2023 | Vietnam & Japan

Top 3 - AI4LIFE (Computer Vision) | Apr 2023 | Da Nang

LANGUAGES & CERTIFICATES

English: TOEIC 720 (ETS, 2023)

Foundation of AI: AI VIETNAM, 2021