

Minh-Thien Nguyen

AI Engineer

📍 Ho Chi Minh City | ✉ minhnnguyent546@gmail.com | 📞 (+84) 373 470 203

in in/minhnnguyent546 | 🌐 minhnnguyent546

EDUCATION

Engineer's degree, Information Technology

08/2021 - 11/2025

Can Tho University

GPA: 3.91/4.0

- **Focus:** Natural Language Processing, Computer Vision.
- Ranked **Top 3 GPA** among Information Technology majors.
- **Thesis:** Building a Smart Enrollment Advisory System Based on an Automatic Dialogue Model.
- **Advisor:** Assoc. Prof. Dr. Nguyen-Khang Pham · **Thesis Grade:** 9.6/10.

EXPERIENCE

AI Engineer

06/2025 - 08/2025

CT Group Corporation

- Conducted research on various large language models such as **ChatGPT**, **Grok**, **Gemini**, and AI Agents like **ManusAI**, OpenManus, and AgenticSeek to integrate and enhance existing systems.
- Developed a chatbot using **RAG** and **FastAPI** to evaluate candidate information and recommend departments for interview candidates.

AI Engineer Intern

12/2024 - 04/2025

TMA Solutions

- Conducted extensive research on 3D human body modeling, facial expression analysis, and relevant AI methodologies to develop and refine the final solution for the required tasks.
- Developed a model to generate a 3D human body from the given height and weight of a user by leveraging **SMPL** and **SMPL-X** models and utilizing the Open3D library for computing parameters and rendering the model.
- Built a **facial expression analysis** module for autistic children as part of the AI system for autism.

PUBLICATIONS

Quoc-Khang Tran, **Minh-Thien Nguyen**, and Nguyen-Khang Pham. "Leveraging Model Soups to Classify Intangible Cultural Heritage Images from the Mekong Delta." *Journal on Information Technologies & Communications (ICT Research)*. Proceedings of the International Conference on Intelligent Systems and Data Science (ISDS), Oct 2025 ([arXiv](#)).

- Proposed a robust framework that integrates the hybrid **CoAtNet** architecture with **model soups** to classify Intangible Cultural Heritage (ICH) images from the Mekong Delta.
- Achieved state-of-the-art results on the ICH-17 dataset (7,406 images across 17 classes) with **72.36%** top-1 accuracy and **69.28%** macro F1-score, outperforming strong baselines (ResNet-50, DenseNet-121, ViT) and previous studies on the ICH-17 dataset.

PUBLICATIONS UNDER REVIEW

Quoc-Khang Tran, **Minh-Thien Nguyen**, and Nguyen-Khang Pham. "ViCLIP-OT: The First Foundation Vision-Language Model for Vietnamese Image-Text Retrieval with Optimal Transport." Preprint submitted to *International Journal of Multimedia Information Retrieval*, 2026. ([arXiv](#)).

- Introduced **ViCLIP-OT**, the **first large-scale vision-language model** specifically developed for **Vietnamese image-text retrieval**, which combines CLIP-style contrastive learning with a novel **Similarity-Graph Regularized Optimal Transport (SIGROT)** loss.
- Achieved state-of-the-art performance on three Vietnamese image-text retrieval benchmarks (UIT-OpenViIC,

KTVIC, and Crossmodal-3600), demonstrating strong zero-shot generalization capabilities.

- Demonstrated that the proposed SIGROT loss effectively captures the global relational structure among samples to improve retrieval accuracy, **reduce the modality gap** between image and text embeddings, and promote better cross-modal alignment in the shared latent space.

PROJECTS

Multiple-Choice Question Answering (MCQA) Pipeline for Complex Technical Documents ([github](#)) 09/2025 - 11/2025

- Developed an end-to-end pipeline to solve MCQA tasks on complex technical documents (PDFs with complex tables, charts, and images).
- Implemented a multimodal extraction layer using **DeepSeek-OCR** for high-fidelity text/table parsing and **Qwen3-VL-4B-Thinking** for automated metadata and structural analysis.
- Developed an advanced **RAG** system integrating Hybrid Search (Dense & Sparse), Query Expansion, and Neighbor Expansion to significantly boost retrieval efficiency. Optimized reasoning and answer generation using **Qwen3-4B-Thinking**, leveraging its thinking capabilities to handle complex reasoning and computations.
- Optimized large-scale inference using **vLLM**, improving throughput and reducing processing time by **4x** compare to using Hugging Face inference.
- **Tech stack:** RAG, DeepSeek-OCR, Qwen3, Qwen3-VL, vLLM, Qdrant, MinerU, Pydantic.

Thesis - Building a smart enrollment advisory system based on an automatic dialogue model ([github](#)) 05/2025 - 08/2025

- Built A Smart Enrollment Advisory System (SEAS) — a RAG-based chatbot for Can Tho University. Backend implemented with async FastAPI and SQLAlchemy 2.0, persisting users, roles, cookie-based authentication, chat sessions, messages, and feedback in Postgres.
- RAG module uses LangChain and Qdrant. Query expansion and reranking are applied to enhance RAG performance. The retrieval component achieved its best results with **0.7713** on Recall@10, **0.7008** on MRR@10, **0.6906** on MAP@10, and **0.7133** on NDCG@10; the average response time is **4.741 seconds**; supports **token streaming** and fast LLM responses via Gemini 2.5 Flash.
- Frontend built with React and TypeScript using TanStack Query/Router, Tailwind, and i18n with **streaming UX**; entire stack containerized and deployed with **Docker Compose**.
- **Tech stack:** FastAPI, LangChain, Qdrant, SQLAlchemy, Pydantic, AWS (Aurora, EC2), ReactJS, TypeScript, Docker, Docker Compose.
- **Advisor:** Assoc. Prof. Dr. Nguyen-Khang Pham · **Thesis Grade:** 9.6/10.

Medical Llama-2 ([github](#)) 08/2024 - 11/2024

- Built a medical chatbot using the **LLaMA 2** model focusing on addressing medical issues and health-related concerns from the user.
- Learned to apply large language models to a domain-specific task, how to fine-tune these large models using parameter-efficient fine-tuning (**PEFT**), specifically **LoRA** and **QLoRA**, and distributed data parallel (**DDP**).
- **Tech stack:** Transformers, Peft, Bitsandbytes, PyTorch, Gradio, LLama.cpp, Docker.

Pre-training GPT-2 ([github](#)) 05/2024 - 08/2024

- Get acquainted with pre-trained and large language models by training the **GPT-2** model from scratch.
- Learned how to scale the training by leveraging the power of **multi-GPU/TPU** training with distributed data parallel via **PyTorch/XLA**.
- Trained the 124M-parameter model on a **TPUv3-8** using the **Fineweb-Edu 10BT** subset, achieving a perplexity of **~21.02**, which outperformed OpenAI's original GPT-2 checkpoints.
- **Tech stack:** PyTorch, Pytorch/XLA for XLA devices (TPU and GPU), Distributed Data Parallel training (DDP).

SKILLS

- **Programming languages:** Python, C/C++.
- **Frameworks:** PyTorch, PyTorch/XLA, TensorFlow, Keras, Transformers, Scikit-learn, LangChain, LangGraph, FastAPI.
- **Model serving:** vLLM, SGLang, TensorRT-LLM.
- **Tools & Technologies:** Git, Docker, Docker Compose, Postman, Llama.cpp, Gradio, Linux, Bash Script, PostgreSQL, Qdrant, OpenAI, GoogleGenAI.
- **Soft skills:** problem-solving, analytical thinking, teamwork, attention to detail.

AWARDS

- Third prize at the 33rd Vietnam Student Olympiad in Informatics (Super Cup division), 2024.
- Third prize at the 2023 ICPC Vietnam National Programming Contest.
- Second prize at the 31st Vietnam Student Olympiad in Informatics (Specialized Informatics division), 2022.
- Third prize at the 2022 ICPC Vietnam National Programming Contest.
- Third prize at the 30th Vietnam Student Olympiad in Informatics (Non-specialized Informatics division), 2021.

CERTIFICATIONS

TOEIC (L&R): 845/990 (issued by the **Educational Testing Service (ETS)** on **June 2024**)

COMMUNITY ENGAGEMENT

- Author of the article **Virtual tree/Cây ảo** - VNOI Magazine, 2024 ([magazine](#)).
- Author of the article **Subtle Techniques with the Xor Operation/Kỹ thuật tinh tế về phép Xor** - VNOI Magazine, 2023 ([magazine](#)).