An Cao

📞 +1-4162549939   ✉ an.cao.link@gmail.com   in ancaoai   ⌗ CA2528357431   G An Cao   🌐 MyWebsite

EDUCATION

University of Toronto　　　　　　　　　　　　　　　　　　　　**Sep 2024 – Dec 2025**
Master of Science in Applied Computing (AI concentration) - GPA: 3.93/4.0　　Toronto, Canada
- Recipient of Vector Scholarship in Artificial Intelligence for top 100 AI master's students in Ontario

Huazhong University of Science and Technology　　　　　　　　**Sep 2020 – Jun 2024**
Bachelor of Engineering in Software Engineering (AI track) - GPA: 3.96/4.0　　Wuhan, China
- Excellent Undergraduate Graduates, Merit Students, Outstanding Undergraduates in Academic Performance

TECHNICAL SKILLS

- **Language & Frameworks**: Python, Pytorch, HuggingFace, DeepSpeed, Langchain, LangGraph, OpenAI, Vertex AI
- **ML Techniques**: Multi-Modal AI, Vision-Language Model, Audio-Language Model, Image Segmentation, Parallel Training, Model Finetuning, Model Deployment, Multi-Agent System, AI Agent, RAG, Embedding Model, Vector Database, Explainable AI, Data Augmentation, Computer Vision, NLP
- **Models**: LLaMa, LLaVa, Qwen, Transformer, BERT, LoRA, SAM, Unet, CLIP, Diffusion, ViT, Audio Flamingo
- **Tools & Infra**: GCP, Azure, AWS, SQL, MongoDB, PostgreSQL, Docker, FastAPI, Flask, Django, MLflow, DVC

EXPERIENCE

Modiface　　　　　　　　　　　　　　　　　　　　　　　　**May 2025 – Present**
Machine Learning Intern　　　　　　　　　　　　　　　　　　　Toronto, Canada
Project: *Digital Dermatologist: Foundational Explainable Vision-Language Model for Skin Health*
- Employed LLMs and Unet for data augmentation, expending dataset by 15 times and imputing 2 absent modalities
- Merged SAM with LLaMa to deliver insights paired with segmentation masks for visually explainable dermatology
- Aligned VLM and segmentation outputs to deliver traceable visual–textual evidence for dermatology insights
- Improved medical analysis accuracy and coverage by 31% over the previous best model through LoRA finetuning
- Boosted skin concern segmentation performance by 23% on IoU above specialized segmentation models
- Applied semi-supervised learning to exploit partial and modality-incomplete data for skin health tasks
- Utilized contrastive learning to finetune an embedding model, enabling chat-based product recommendations

Vector Institute　　　　　　　　　　　　　　　　　　　　　**June 2025 – Present**
AI Technical Specialist　　　　　　　　　　　　　　　　　　　Toronto, Canada
- Integrated Qwen multi-modal LLM into audio-text RAG system for real-time voice-based grounded health insights
- Led a RAG system optimized for live tabular stock data, matching human performance with 85% retrieval recall
- Designed a multi-agent system for Anti-Money Laundering, attaining 76% accuracy in detecting info inconsistencies

Vector Institute　　　　　　　　　　　　　　　　　　　　　**Sep 2024 – May 2025**
Machine Learning Associate　　　　　　　　　　　　　　　　　Toronto, Canada
Project: *DiligenceGPT: AI for Due Diligence*
- Applied multi-modal LLMs to extract finance data from uncurated & unstructured documents with 89% coverage
- Built a real-time RAG agent with vector search and minute-level live data, providing traceable response generation
- Designed a LLM + Neural Network quantitative evaluator with 97% consistency, exceeding human expert performance
- Orchestrated async model deployment to parallelize inference, reducing response latency by 70% on average

Project: *Audience Builder: Conversational Database Agent for Synthetic Society*
- Created a conversational RAG agent that suggests relevant database values, accelerating complex query formulation
- Leveraged LLMs to decompose complex requests into atomic sub-queries, ensuring retrieval covers all specified objects
- Solved semantic retrieval bottlenecks by fine-tuning the embedding model, achieving a 49% improvement in recall
- Implemented an async streaming backend with FastAPI to maintain low-latency responses under concurrent user loads

Huazhong University of Science and Technology　　　　　　　　**Sep 2021 – Aug 2024**
Deep Learning Research Assistant　　　　　　　　　　　　　　　Wuhan, China
- Built Diff-STAR, a Student-Teacher model combining Diffusion and ViT, achieving SOTA in image harmonization
- Proposed LisaCLIP, a zero-shot text-driven adaptive model, enabling image manipulation using text without training
- Collaborated on Virtual Try-On and Street Semantic Segmentation projects, refining models and doing ablation studies

PUBLICATIONS

Diff-STAR (**First Author, Published by IMAVIS, SCI Q1**)　　　　　**Sep 2023 – Aug 2024**
A. Cao and G. Shen, "Diff-STAR: Exploring student-teacher adaptive reconstruction through diffusion-based generation for image harmonization," Image Vis. Comput., vol. 151, p. 105254, Nov. 2024, doi: 10.1016/j.imavis.2024.105254.

LisaCLIP (**First Author, Accepted by IJCNN as ORAL**)　　　　　**Sep 2022 – Feb 2023**
A. Cao, Y. Zhou, and G. Shen, "LisaCLIP: Locally Incremental Semantics Adaptation towards Zero-shot Text-driven Image Synthesis," in 2023 International Joint Conference on Neural Networks (IJCNN), Jun. 2023, pp. 1–10. doi: 10.1109/IJCNN54540.2023.10191516.