

Trung Dao

tdao6@wisc.edu | [linkedin.com/in/trung-dt880](https://www.linkedin.com/in/trung-dt880) | github.com/trungdt880

RESEARCH INTERESTS

My research centers on Vision-Language Models (VLMs) and Vision-Language-Action (VLA) systems, building agents that perceive physical environments, reason about humans and objects, and act in the real world. I bring a track record in efficient generative modeling (diffusion distillation, one-step generation) and real-time, on-device perception (autonomous-driving vision, always-on VLMs), and aim to translate few-step generative modeling and edge inference into VLA policies and small world action models that run in real time on embodied hardware.

EDUCATION

University of Wisconsin-Madison

Ph.D. in Computer Science (Advisor: Prof. *Yong Jae Lee*)

USA

Jan 2026 – Expected 2030

Thang Long University

Bachelor of Computer Science; GPA: 9.0/10.0 (*Valedictorian*)

Vietnam

Aug 2016 - Apr 2021

EXPERIENCE

University of Wisconsin-Madison

Graduate Research and Teaching Assistant

USA

Jan 2026 - Present

- **Research:** VLA policies and world models for embodied agents, with a focus on inference-efficient (distilled / few-step) action generation for real-time, on-robot deployment.
- **Teaching:** Teaching Assistant for *Introduction to Computer Engineering* (Spring 2026).

Qualcomm AI Research

Staff Machine Learning Engineer

Vietnam

Nov 2024 - Jan 2026

- Fast-tracked promotion to **Staff Engineer** after a “Far Exceeds Expectations” (5/5) rating in the 2025 annual review.
- **On-Device Agentic AI:** Led the end-to-end quantization and deployment of InternVL3.5 (VLM) for always-on mobile screen understanding on Snapdragon — fine-tuned, quantized, and shipped a performant on-device inference flow under tight power/latency budgets with < 5% accuracy degradation.
- **Frontier Model Deployment:** Built the W4A16 quantization and deployment of LFM-2, the first hybrid transformer model to run on the Snapdragon Gen 5 chip, reaching **9000 tok/s** prefill and **90 tok/s** decode (4K context). Quantized and deployed further frontier models (Nemotron, InternVL) for resource-constrained edge devices.
- **Advanced Quantization:** Resolved critical W4A16 accuracy collapse in lightweight VLMs caused by activation/weight outliers, applying techniques (rotation, SpinQuant, AdaScale, GPTQ) selected per model and deployment to recover high performance on Snapdragon devices.
- **Efficient Diffusion Models:** Introduced a block-based distillation for FLUX.1 (40% smaller, > 95% of HPSv2 retained); developed an attention-based text-encoder distillation enabling a T5-XXL → T5-base swap in PixArt- α with < 2% HPSv2 drop.
- Joined via MovianAI; transitioned to Qualcomm after its acquisition (Apr 2025). *Past role: Senior Machine Learning Engineer.*

VinAI Research

Research Resident

Vietnam

Mar 2023 - Oct 2024

- **Advisor:** Dr. *Anh Tran*, Dr. *Cuong Pham*.
- **Research Focus:** Generative vision models, with an emphasis on GANs and diffusion models.
- **Selected works:**
 - Improved the quality of one-step and few-step text-to-image diffusion models [**P2**, **P4**, **P5**].
 - Introduced a novel diffusion architecture integrating Mamba for improved efficiency and scalability [**P3**].
 - Built a large-scale extreme-pose face dataset to improve synthesis quality and benchmark face recognition [**P1**].
- **HPC cluster management:** Managed and optimized a $48 \times$ A100 GPU cluster, raising mean real-time GPU utilization from **15–20%** to **83%** via a novel preemptive queuing strategy.

VinAI Research

AI Engineer

Vietnam

Dec 2020 - Mar 2023

- **Advisor:** Dr. *Dzung Nguyen*, Dr. *Anh Tran*, Prof. *Minh Hoai Nguyen*.
- **Face Recognition Module** (Role: Module Owner)
 - Built multi-node training (up to 60M images) and a SLURM framework for profiling, tuning, and optimization; shipped models for masked access control and surveillance CCTV at **50K daily active identities**.

- Achieved **8th** overall (**2nd** on Masked Dataset) at the ICCV21-MFR Competition, Jul 2022.
- Quantized and deployed a 3-model module on Qualcomm AIC100 (up to 30 concurrent streams), plus NVIDIA (TensorRT) and Android (ONNX, MNN, NCNN); built supporting tools for visualization, video inference, and semi-automated data cleaning.
- **Face Detection Module** (Role: Module Co-owner)
 - Trained a multi-task masked-face detector for surveillance cameras, handling tiny faces, blocking artifacts, and occlusions; generated pseudo-masks via 2D/3D methods.
 - Co-built the AI SDK: optimized and deployed models on Xilinx with an asynchronous multi-stream (DeepStream) flow running up to **60 streams** on Xilinx ZCU104.
- **Traffic Sign/Light Recognition Module for Autonomous Driving** (Role: Module Co-owner)
 - Designed a CVAT-based pipeline to accelerate video labeling, yielding a dataset of *six superclasses and 317 child classes*; co-managed the labeling team for quality.
 - Built real-time perception for a moving autonomous-driving platform: a hierarchical multi-task recognizer (macro-F1 **98.3** on a private long-tailed 171-class set) with a ReID tracker for lighting robustness, deployed on-vehicle via TensorRT — the latency and robustness constraints embodied agents share.
- **Other projects**
 - **Noise Cancelling on Smartphone** Responsible for converting models across various frameworks (PyTorch, TensorFlow, ONNX) into TFLite, followed by quantization and smartphone deployment. Optimized existing algorithm with FFT, achieving a **40%** runtime reduction.
 - **SmartData** Redesigned the data labeling pipeline of the backend system built with Flask. Introduced a new end-to-end multi-step labeling feature, improving labeling efficiency by **30%**.

PROFESSIONAL SERVICES

Reviewer: CVPR(2023, 2024, 2025, 2026), NeurIPS(2024, 2025), ECCV(2024, 2026), ICCV(2025), ICLR(2025), WACV(2025), ACCV(2022, 2024).

Outstanding Reviewer: CVPR 2026 (Top 5%).

CERTIFICATES, HONORS AND AWARDS

Top 5 Exceptional Vietnamese AI Talents

VNExpress AI Awards 2025

Rank 2nd (Masked Dataset); 8th (Overall)

ICCV21-MFR Competition 2021

Academic Excellence Scholarship

Thang Long University 2016-2021

First Runner-up

VietAI Machine Learning Foundation Hanoi 2020

First Runner-up

Fintech Track, Junction X Hanoi 2018

Rank 76th

ICPC Asia Hanoi Regional Contest 2018

SKILLS SUMMARY

Languages: C++, Python, Unix scripting, SQL

Tools: PyTorch, TensorFlow, TensorRT, AIMET, ONNX, NCNN, MNN, OpenCV, Docker, Git, Jira

REFERENCES

Prof. Yong Jae Lee: Susan Beth Horwitz Professor, University of Wisconsin-Madison: yongjaelee@cs.wisc.edu

Dr. Dung Nguyen: Director, Engineering, Qualcomm AI Research, Vietnam: dungnt@qti.qualcomm.com

Dr. Anh Tran: Principal Engineer, Qualcomm AI Research, Vietnam: anhtra@qti.qualcomm.com

Prof. Minh Hoai Nguyen: Deputy Director of AIML, University of Adelaide, Australia: mh.nguyen@adelaide.edu.au

PUBLICATIONS

(*) denotes equal contribution.

- [P1] **Trung Dao***, Duc Hong Vu*, Cuong Pham and Anh Tran. "EFHQ: Multi-purpose ExtremePose-Face-HQ dataset." CVPR, 2024.
- [P2] **Trung Dao**, Thuan Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, Anh Tran. "SwiftBrushV2: Make Your One-step Diffusion Model Better Than Its Teacher." ECCV, 2025.
- [P3] Hao Phung*, Quan Dao*, **Trung Dao**, Hoang Phan, Dimitris N. Metaxas, Anh Tran. "DiMSUM: Diffusion Mamba - A Scalable and Unified Spatial-Frequency Method For Image Generation." NeurIPS, 2024.
- [P4] Quan Dao*, Hao Phung*, **Trung Dao**, Dimitris N. Metaxas, Anh Tran. "Self-Corrected Flow Distillation for Consistent One-Step and Few-Step Image Generation." AAAI, 2025.
- [P5] Viet Nguyen*, Viet Nguyen*, **Trung Dao**, Toan Tran, Anh Tran. "SNOOPI: Supercharged One-step Diffusion Distillation with Proper Guidance." ICCV, 2025.
- [P6] Anh Nguyen*, Viet Nguyen*, Duc Vu, **Trung Dao**, Chi Tran, Toan Tran, Anh Tran. "Improved Training Technique for Shortcut Models." NeurIPS, 2025.