Hoai-Chau Tran

linkedin.com/in/hoai-chau-tran
chauht2@illinois.edu
hchautran.github.io
217-841-0396

EDUCATION

University of Illinois at Urbana-Champaign (UIUC) — Illinois, USA
Ph.D. Student (Expected day of graduation: 5/2030) — *8/2025-present*

- **Research interest:** Optimization for efficient 3D Transformers and multimodal models including (VLM, VLA), kernel development for efficient attention mechanisms, quantization and compression techniques, long-range sequence learning, and embodied AI systems.
- **GPA:** 4.0/4.0.

University of Science, VNU-HCM — Ho Chi Minh City, Vietnam
Bachelor of Science, Advanced Program in Computer Science — *2018-2022*

- **GPA:** 8.9/10.0.

SELECTED PUBLICATIONS

List of publications on Google Scholar.

1. Tuan Anh Tran, Duy Minh Ho Nguyen, **Hoai-Chau Tran**, Michael Barz, Khoa D Doan, Roger Wattenhofer, Vien Anh Ngo, Mathias Niepert, Daniel Sonntag, Paul Swoboda. How Many Tokens Do 3D Point Cloud Transformer Architectures Really Need?. *Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).*.
2. **Hoai-Chau Tran**, Duy M. H. Nguyen, Duy M. Nguyen, Trung-Tin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Y. Zou, Binh T. Nguyen, Mathias Niepert. Accelerating Transformers with Spectrum Preserving Token Merging. *Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024).*.
3. **Hoai-Chau Tran**, Anh-Duy Tran, Kim-Hung Le. DetectVul: A statement-level code vulnerability detection for Python. *Future Generation Computer Systems (FGCS). 163 (2025) p. 107504. 2025*.
4. **Hoai-Chau Tran**, Chi H Nguyen, Duy Minh Ho Nguyen, Mathias Niepert, Fan Lai, Khoa D Doan SparseSAM: Structured Sparsification of Attention in Segment Anything Models. *In Submission*.

RESEARCH EXPERIENCE AND PROJECTS

SparseSAM — Illinois, USA
Research Assistant — *8/2025-present*

- **Description:** Developing *SparseSAM*, a sparse attention mechanism for Segment Anything Models (SAM, SAM2). SparseSAM reduces the latency of the attention operator by *up to* $\sim 9\times$ *for global attention and* $7\times$ *for local attention* while maintaining performance comparable to dense attention. This work is currently under submission.
- **Advisors:** Prof. Khoa D. Doan, Prof. Fan Lai

MyGPT2: Custom CUDA Implementation for LLM Inference — Illinois, USA
Independent Research — *9/2025-11/2025*

- Implemented GPT-2 inference entirely in *custom CUDA kernels* (attention, matrix multiplication, *quantization*) from scratch without relying on high-level frameworks like PyTorch or cuBLAS.
- Designed and optimized *FlashAttention kernels* using advanced techniques including shared-memory tiling, Split-K decomposition, local attention windows, and customized Tensor Core instructions, achieving 75% throughput and memory-efficiency compared to the flash-attn package.
- Integrated *Activation-Aware Quantization (AWQ)* with calibration, per-channel scaling, and inference-time dequantization to substantially reduce memory bandwidth usage while preserving model accuracy.

VinUni-Illinois Smart Health Center, VinUniversity — Ha Noi, Vietnam
Research Assistant — *8/2024-7/2025*

- **Description:** Research on *3D Transformer architectures* (PTv3, Sonata, etc.) and development of *GitMerge3D*, a token compressing algorithm for *accelerating 3D point cloud transformers*. The method estimates voxel token importance from spatial structures and selectively merges low-importance tokens, enabling up to 90% token reduction while preserving accuracy. Achieved substantial efficiency gains by reducing PTv3 computation from 107.5 GFLOPs to 19.9 GFLOPs and memory footprint from 10.12 GB to 1.6 GB, with negligible mIoU degradation (e.g., 77.6 $\rightarrow$ 77.4).

IEC Research Group, University of Information Technology, VNU-HCM Ho Chi Minh, Vietnam
Research Assistant *1/2023-7/2024*

- **Description:** Collect data and developed DetectVul—first statement-level Python vulnerability detection model without graph extraction; achieved +25.45% F1 improvement over GCN and +18.05% over GAT across 211K statements. Released full pipeline, datasets, and model artifacts for reproducibility.
- **Collaborators:** Anh-Duy Tran, Kim Hung Le
- **Advisor:** Prof. Kim Hung Le

INDUSTRY EXPERIENCE

MoMo e-wallet Ho Chi Minh, Vietnam
Software Engineer *8/2021-12/2022*

- Developed and maintained REST APIs for personal finance management, including budgets, expenses, and transaction histories.
- Created tools and scripts to assist other teams in managing development databases and benchmarking services's performance and resilience.
- Integrated SQLite for local data storage to enable faster transaction history query performance on mobile devices.

SKILLS

1. **Programming languages:** Python, C/C++, CUDA, Java, SQL.
2. **Technologies:** PyTorch, CUDA, Numpy, HuggingFace, Git, Scikit-learn, AWS, Docker, Linux, LaTeX.
3. **Languages:** English *(fluent)*, Vietnamese *(native)*.

RELEVANT COURSES

- **CS598 - System For Gen AI** by Prof. Fan Lai *(UIUC - Spring 2026)*
- **CS445 - Computational Photography** by Prof. Yuxiong Wang *(UIUC - Spring 2026)*
- **CS446 - Machine Learning** by Prof. Liangyan Gui *(UIUC - Fall 2025)*
- **CS483 - Applied Parallel Computing** by Prof. Volodymyr Kindratenko *(UIUC - Fall 2025)*
- **Deep Learning Specialization** by Prof. Andrew Ng *(Coursera)*
- **Linear Algebra** by Prof. Nguyen Huu Anh *(University of Science - Summer 2019)*
- **Probability and Statistics** * by Prof. Luu Quoc Toan *(University of Science - Fall 2020)*
- **Calculus I,II,III** by Prof. Le Van Luyen *(University of Science - Fall 2019, Spring 2020, Fall 2021)*

PROFESSIONAL SERVICES

1. Reviewer: NeurIPS, ICLR, ICML
2. Volunteer: NeurIPS

REFERENCE

1. **Prof. Fan Lai**, Assistant Professor at the Siebel School of Computing and Data Science - University of Illinois at Urbana-Champaign (UIUC), USA *fanlai@illinois.edu*
2. **Prof. Khoa D Doan**, Associate Director at the VinUni-Illinois Smart Health Center. Assistant Prof. at College of Engineering & Computer Science, VinUniversity, Vietnam *khoa.dd@vinuni.edu.vn*

*Obtained A+ for excellent performance.