Jui-Chien (Eric) Tsou

☎ (+886) 906-819-625 ✉ tsouoeric@gmail.com Jui-Chien Tsou 🌐 Website

Education

National Taiwan University (NTU) *Taipei, Taiwan*

B.S. in Computer Science, College of Electrical Engineering and Computer Science *Sep 2022 – Jun 2026*

Overall GPA: 4.12 / 4.3, Last 60: 4.21 / 4.3

Publications

Boosting Small Object Tracking via Collaborative Detection Transformer

International Conference on Machine Vision and Applications (MVA) **Oral**, 2025

- Analyzed hybrid parallelism on multi-GPU cluster by examining NVLink bandwidth utilization and GPU hardware utilization, identifying communication bottlenecks and impact on scalability

Instruction-Tuned LLMs for Multilingual Medical ASR and Privacy Entity Extraction

The 20th World Congress on Medical and Health Informatics (MEDINFO) **Workshop**, 2025

- Analyzed hybrid parallelism on multi-GPU cluster by examining NVLink bandwidth utilization and GPU hardware utilization, identifying communication bottlenecks and impact on scalability

MVA 2025 Small Multi-Object Tracking for Spotting Birds Challenge

International Conference on Machine Vision and Applications (MVA), 2025

- Analyzed hybrid parallelism on multi-GPU cluster by examining NVLink bandwidth utilization and GPU hardware utilization, identifying communication bottlenecks and impact on scalability

Research Experience

Accelerating Optical Character Recognition Models *Taipei, Taiwan*

Undergraduate Research, Advisor: Prof. Chun-Yi Lee, Dept. of CSIE, NTU *Dec 2025 – Present*

- Analyzed hybrid parallelism on multi-GPU cluster by examining NVLink bandwidth utilization and GPU hardware utilization, identifying communication bottlenecks and impact on scalability
- Applied 8-bit quantization to parameters, activations, and KV cache, and leveraged kernel fusion to reduce external data access, achieving 15× inference throughput improvement for Transformer models

Distributed LLM Inference across Heterogeneous CPU and GPU Platforms *Taipei, Taiwan*

Undergraduate Research, Advisor: Prof. Chun-Yi Lee, Dept. of CSIE, NTU *Feb 2025 – Dec 2025*

- Applied 8-bit quantization to parameters, activations, and KV cache, and leveraged kernel fusion to reduce external data access, achieving 15× inference throughput improvement for Transformer models
- Leveraged AVX-512 and AMX-bf16 on Intel Xeon CPUs to accelerate matrix operations, and used NUMA binding to colocate threads and memory, achieving 3× inference latency speedup for Transformer models

Performance Analysis of Parallel Influence Maximization Algorithms on High Performance Computing Systems *Taipei, Taiwan* *Sep 2025 – Dec 2025*

Undergraduate Research, Advisor: Prof. Chun-Yi Lee, Dept. of CSIE, NTU

- Analyzed efficient dataflow for matrix multiplication (GEMM), vector operations (GEMV), and special functions in Transformer models to maximize on-chip memory data reuse and external HBM bandwidth utilization
- Developed simulation framework using SystemVerilog and C++ via DPI to model HBM behavior at RTL, enabling billion-parameter simulations with 20× speedup for verification and debugging

Working Experience

Simulation and Development of Vision Language Action Model on Humanoid Robots *Taoyuan, Taiwan* *Jul 2024 – Aug 2024*

Cloud Application Researcher, Quanta Cloud Technology

- Applied Gabor transform for fine-grained spectral observation and dynamic masking with frequency-domain Wiener filtering to recover noiseless signal components
- Proposed time-variant noise amplitude estimation method by averaging energy across masked frequency

regions, capturing dynamic noise patterns and adapting filtering to improve audio quality

Agentic LLM with Autonomous Reasoning, Planning, and Tool Interactions — *Taipei, Taiwan*

Research Assistant, Mentor: Prof. Yuh-Jye Lee, Academia Sinica — *Jul 2024 – Aug 2024*

- Applied Gabor transform for fine-grained spectral observation and dynamic masking with frequency-domain Wiener filtering to recover noiseless signal components
- Proposed time-variant noise amplitude estimation method by averaging energy across masked frequency regions, capturing dynamic noise patterns and adapting filtering to improve audio quality

Competitions and Awards

2025 SC Student Cluster Competition (SCC25) — *St. Louis, MO, USA*

Overall Winner — *Nov 2025*

- Achieved victory in an international contest featuring teams from UCSD, ETH Zurich, and Nanyang Technological University, etc. (In total of 8 teams).
- Led on-site cluster deployment and implemented multi-node MLPerf Inference (Llama2-70B), including system optimization, CPU/GPU profiling, and performance tuning.

2025 APAC HPC-AI Competition (HPC-AI) — *Osaka, Japan*

Second Place & Best AI Performance Award — *Oct 2025*

- Secure second place among **48 international teams**, competing against prestigious institutions like National Tsing Hua University and Nanyang Technological University.
- Reduced communication time by **7×** and achieved **2×** improvement in offline inference throughput for DeepSeek-R1.

2025 ASC Supercomputer Challenge (ASC25) — *Qinghai, China*

First Prize & Group Competition Award — *May 2025*

- Earned first prize and group competition award in an international contest with **118 teams**, including strong competitors like Peking University, Tsinghua University, and Shanghai Jiao Tong University.
- Build and optimized clusters and parallelized DNA sequence alignment execution pipeline & HPL, HPCG tuning.

2025 Small Object Tracking Challenge (MVA) — *Kyoto, Japan*

Third Place — *Jul 2025*

- Earned third place in a small object tracking challenge at a international conference.
- Fine-tuned image recognition models with collaborative hybrid assignments and tracking with slicing-aided hyper inference.

2025 National Intercollegiate Artificial Intelligence Competition (AICUP) — *Taipei, Taiwan*

Honorable Mention — *Jun 2025*

- Competed nationally with thousands of teams, secured 6th place among many teams from National Taiwan University and National Tsing Hua University.
- Combining LLMs, instruction tuning, and hybrid post-processing for medical speech privacy protection.

Teaching Experience

CSIE5213 Parallel Programming, Fall 2025 — *Sep 2025 – Dec 2025*

Teaching Assistant, Dept. of CSIE, NTU

- Led an assignment and instructed students on the use of high-performance computing servers.

Knowledge & Skills

Software Languages:

- Python (PyTorch, TensorFlow, Numpy, Pandas)
- C / C++ (Pthread, OpenMP, OpenMPI), GPU Programming (CUDA, ROCm)
- Open-Source LLM Serving Frameworks: TensorRT-LLM, vLLM, SGLang, and llama.cpp