

Wen-Chi Tsai

✉ vickytsa@andrew.cmu.edu 📞 +1 (412) 224-7224 🌐 Portfolio 🔄 vicky0619 📄 wen-chi-tsai

WORK EXPERIENCE

Software Engineering Intern Jun. 2025 – Mar. 2026
ASUS, Open Cloud Infrastructure Software R&D Team Taipei, Taiwan

- Cut a full regression pass from **~16 hours to ~1–2 hours** with a **TypeScript end-to-end (E2E) test suite** covering ~95% of previously manual cases across Linux, Windows, and macOS
- Automated issue triage with a **GitLab Model Context Protocol (MCP) agent**, pairing an LLM classification prompt with MCP tool calls to auto-classify, label, and route issues
- Shipped four **context providers into a VS Code AI coding assistant**, adding commit-history, OS-metadata, current-file, and repository-map context selected from an evaluation of candidates
- Cut hands-on onboarding support by authoring the assistant's **Docusaurus documentation site**

Startup Engineer Jun. 2026 – Aug. 2026
RoomMuse Pittsburgh, PA

- Caught unbuildable layouts at design time rather than on install day with a **six-rule geometric conflict engine** that blocks any unit sealing a socket, pipe, or doorway while passing intended connections (**45 tests**)
- Removed manual fixture surveying by localizing sockets from the same **Swift and RealityKit** scan with a **classical-CV pipeline** (thresholding, connected components, NMS) that ray-casts detections onto wall planes

Research Assistant, Edge LLM Compression Sep. 2025 – Feb. 2026
Wireless Ad-Hoc and Sensor Networks Laboratory, NCU Taoyuan, Taiwan

- Sharded teacher-student distillation across 3 GPUs with **DeepSpeed ZeRO-2**, splitting optimizer state and gradients, since **ZeRO-3 sharding corrupts GPTQ-packed quantized weights**
- Recovered **4-bit accuracy at 2-bit model size** (IoU 0.9998, 1.18 GB) by fine-tuning a Llama-3.2 student with **GPTQ quantization and LLM-Neo knowledge distillation** in PyTorch
- Surveyed **llama.cpp, vLLM, and SGLang** for edge deployment, mapping **latency versus throughput** across PagedAttention, continuous batching, and prefix-cache reuse

SELECTED PROJECTS

Dual-Strategy UAV Path Planning Sep. 2024 – Dec. 2025

- Designed a dual-strategy planner trading **layered MST-DFS** for structured-space traversal against **collision-aware A* replanning**, cutting path-generation time over $5\times$ and path length up to 14%
- Built the **3D point-cloud pipeline in C++ and Python**, slicing raw .pcd scans into layers and chaining **per-layer KNN and MST** into a DFS traversal, with normal-vector correction for full coverage
- Validated **100% collision-free operation in AirSim** against a random-order baseline on real-world building models

Museum Artifact Recognition App May 2024 – Nov. 2024

- Shipped **on-device model inference in an iOS app**, running the team's YOLOv8n and Pix2Pix reflection-removal models in **Swift with TensorFlow Lite** at 94.5% accuracy across 7 classes
- Integrated the iOS client with the team's **AWS EC2 REST inference API**, adding a cloud path for multi-user recognition

PUBLICATIONS

Wen-Chi Tsai, Yi-Peng Liu, Yu Wu, Yun-Sheng Chen, Chung-Lun Lai, Yu-Yu Chang, Min-Te Sun. "UAV Path Planning for Building Structure Inspection." In *14th International Conference on Ubi-Media Computing (Ubi-Media 2026)*, Penang, Malaysia.

EDUCATION

Carnegie Mellon University (CMU), School of Computer Science Aug. 2026 – Dec. 2027 (Expected)
Master of Science in Intelligent Information Systems, Language Technologies Institute Pittsburgh, PA

- **Relevant coursework:** Deep Learning Systems, Multimodal Machine Learning, Advanced Natural Language Processing

National Central University (NCU) Sep. 2022 – Jun. 2026
Bachelor of Science in Computer Science and Information Engineering | GPA: 3.99/4.30 Taoyuan, Taiwan

- **Relevant coursework:** Systems Programming, Generative AI (Text & Image), Pattern Recognition, Data Science