John (Jui-Chiang) Wei

(+886) 906539110 | johnwei@access.ee.ntu.edu.tw | LinkedIn | Google Scholar

Education

National Taiwan University (NTU), Taiwan Sep 2025 – Present
M.S. in Graduate Institute of Electronics Engineering, Integrated Circuits and Systems Group
Advisor: Prof. An-Yeu (Andy) Wu

National Taiwan University (NTU), Taiwan Sep 2022 – Jun 2025
B.S. in Electrical Engineering GPA: 3.85/4.0

Professional Experience

Access IC Design Lab, NTU Sep 2025 – Present
Graduate Researcher, advised by Prof. An-Yeu (Andy) Wu

- Research on efficient AI model compression and accelerator-aware optimization for emerging neural architectures.
- Research expertise in post-Transformer models such as Mamba-based and linear attention architectures.

Academia Sinica Jul 2024 – Aug 2024
Research Intern / AI Engineer Intern

- Worked on TinyML, sustainable AI, and efficient model deployment.

Publications

HyQuant: A Unified Quantization Framework for Hybrid Mamba–Transformer Vision Models ICIP 2026
J.-C. Wei, B.-Y. Shi, An-Yeu Wu

- First author. Proposed a systematic framework for post-training quantization of hybrid vision architectures such as MambaVision and VSSD.

Multi-Distillation from Speech and Music Representation Models ASRU 2025
J.-C. Wei, Y.-C. Lin, F. Ritter-Gutierrez, and H.-y. Lee

- First author. Proposed a multi-teacher distillation framework for unifying speech and music representation models.

Distilling a Speech and Music Encoder with Task Arithmetic Interspeech 2025
F. Ritter-Gutierrez, Y.-C. Lin, **J.-C. Wei**, J. H. M. Wong, E. S. Chng, N. F. Chen, and H.-y. Lee

- Co-authored a task-arithmetic-based distillation approach for combining speech and music encoders.

Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks ICLR 2024

- Contributed to a large-scale benchmark for evaluating spoken language models across diverse tasks.

Projects

Multi-Core GEMM Accelerator 2025

- Designed a multi-core GEMM accelerator supporting distributed memory, AXI-based data movement, and controller-managed task scheduling across processing elements.

ViT Patch Embedding Accelerator 2025

- Implemented a Vision Transformer patch embedding accelerator with INT8 quantization support, AXI memory access, and clock-domain-crossing synchronization.

BCH Decoder 2024

- Implemented a BCH decoder supporting hard-decision and soft-decision decoding for multiple code lengths, including Berlekamp-based error-location computation and Chase decoding.

Honors and Awards

- Dean's List / Academic Excellence Award, National Yang Ming Chiao Tung University * 2.
- Calculus Award, National Yang Ming Chiao Tung University, ranked 2nd university-wide, 2021.
- TOEIC Gold Certificate, 910/990.

Languages and Skills

Technical Skills	Python, PyTorch, C/C++, MATLAB, Verilog/SystemVerilog, JavaScript
EDA Tools	Cadence NC-Verilog, Synopsys Design Compiler, Cadence Innovus, Calibre DRC/LVS
Research Areas	Model Compression, Post-Training Quantization, IC Design, AI Accelerators, Vision Models
Languages	Mandarin, English