You-Rui Lin (Nick)

AI Research • AI Infrastructure • Distributed Systems

Location Kaohsiung, Taiwan
Phone +886 960 802 253
Email nicklin910221@gmail.com
LinkedIn linkedin.com/in/you-rui-lin
GitHub github.com/NickLin910221



Profile

M.S. student in AI with hands-on experience in GPU cluster infrastructure and end-to-end computer-vision systems. Current research focuses on KV-cache eviction and compression for efficient LLM inference. Seeking AI research or AI infrastructure roles.

Education

M.S., College of Artificial Intelligence
Jul 2024 – Present

National Yang Ming Chiao Tung University, Taiwan

- **GPA:** 4.02 / 4.30 **Advisors:** Yu-Chee Tseng, Jen-Jee Chen
- **Thesis (in progress):** EntropyKV: Entropy-based KV Cache Eviction and Compression
- **Coursework:** Computer Vision, Distributed Systems, High-Performance Computing, System Architecture Design, Microservices

B.S., Computer Science and Information Engineering
Sep 2020 – Jun 2024

National University of Tainan, Taiwan

- **Advisor:** Tzung-Shi Chen
- **Graduation Project:** Industry-university cooperation license plate recognition system; extended research on generative AI image generation

Publications

- **You-Rui Lin**, Jen-Chieh Wu, Tzu-Ti Wei, Jen-Jee Chen, Yu-Chee Tseng. *DetM-Tracker: Deterministic and Memory-Efficient VIS for Autonomous Industrial Inspection*. arXiv preprint arXiv:2610.29212, 2026.

Projects & Research

Lab AI Infra: Disaggregated Storage & Compute Scheduling
2025 – 2026

NYCU PAIRLab

- Architected a compute-storage disaggregated cluster on Kubernetes and Ceph CSI: 7 nodes, 42 TB NVMe, 26 GPUs (L40S, V100, RTX 6000 Ada).
- Let about 60 researchers request specific GPU models on demand without manual dataset migration.
- Cut manual operations and data migration overhead by over 90%.

PoC: Cable Routing Inspection
2025

NYCU PAIRLab

- Led an end-to-end project from scratch to solve industrial cable routing using computer vision, covering system design, data collection, model training, and deployment.
- Designed the complete microservices architecture and evaluated smart glass hardware through hands-on testing before procurement.
- Trained the model from scratch under NDA while accounting for real-world constraints (compute, power, latency), reaching 60% end-to-end inspection accuracy with limited GPU resources.

Highway License Plate Recognition (LPR)
Nov 2024 – Aug 2025

DETR-based recognition model, in collaboration with a third-party company

- Designed a DETR-based architecture and trained it from scratch, optimized for higher throughput on highway traffic.
- Achieved 98.6% recognition accuracy.

Experience

AI Dataset Intern
Jul 2025 – Aug 2025

Wiwynn Corp.

- Deployed data collection environments on the production line and defined collection workflows with operators for an industrial cable routing project.
- Collected real-world industrial data for end-to-end model training.

AI Labeling Engineer Intern
Jul 2023 – Sep 2023

Asus Cloud Corp.

- Performed large-scale data annotation and quality control; helped refine labeling guidelines to improve consistency.

Skills

Programming Python, C/C++, Linux
AI / Systems PyTorch, vLLM, Kubernetes, Docker, Ceph, Microservices, GPU Cluster Orchestration
Languages Mandarin Chinese (native), English (IELTS 5.5)

References available on request.