Jorge Tyrakowski

Portfolio and demos: **jorgetyrakowski.com**
jorgetyrasc@gmail.com | linkedin.com/in/jorgetyrakowski | github.com/jorgetyrakowski | Google Scholar

Education

National Taiwan University (NTU) — Taipei, Taiwan
M.S. in Computer Science and Information Engineering — *Sep. 2026 – Jun. 2028 (Expected)*

Advisor: Prof. Wen-Huang Cheng | AIMM Research Group, Communications and Multimedia Laboratory (CMLab)
Coursework: Applied Deep Learning, Computer Vision Practice with Deep Learning, Multimodal Artificial Intelligence

National Yang Ming Chiao Tung University (NYCU) — Hsinchu, Taiwan
B.S. in Computer Science — *Sep. 2022 – Jun. 2026*

Relevant Coursework: Visual Recognition using Deep Learning, Introduction to Machine Learning, Artificial Intelligence Capstone, Computer Vision, Computer Vision for UAV Autopilot, Data Mining, Image and Video Generation

Fo Guang University — Yilan, Taiwan
Full-time Mandarin Chinese Program (570 hours) — *Oct. 2021 – Aug. 2022*

Experience

AIMM Research Group, National Taiwan University — Taipei, Taiwan
Graduate Researcher — *Sep. 2026 – Present*

- Researching efficient inference for multimodal LLMs (MLLMs) in an industry-sponsored project: visual token pruning and KV cache optimization
- Focusing on multi-view MLLMs for autonomous driving, where every extra camera adds visual tokens and inference cost

Institut de Robòtica i Informàtica Industrial (IRI), CSIC-UPC — Barcelona, Spain (remote)
Research Collaborator — *Sep. 2026 – Present*

- Extending the dataset of "Urban Risk-Aware Navigation via VQA-Based Event Maps for People with Low Vision"↗ (major revision at IEEE T-ITS): collecting pedestrian sequences and annotating them by risk category
- Running inference with the VQA models benchmarked in the paper (ViLT, LLaVA, InstructBLIP, Qwen-VL) on the new data

Industrial Technology Research Institute (ITRI) — Hsinchu, Taiwan
R&D Intern (AI/ML) — *Jul. 2025 – Jul. 2026*

Taiwan's largest national applied research institute; the R&D organization that spun off TSMC and UMC.

- Built a real-time digital guide for ITRI's museum: a web-based, bilingual (English/Chinese) conversational avatar that answers visitors' questions from a RAG knowledge base and adapts its tone to the person on camera, with the LLM, ASR, VLM and TTS all running on one NVIDIA Jetson AGX Thor
- Optimized LLM inference for edge deployment through model selection, NVFP4 quantization trade-offs and pipeline restructuring to cut end-to-end latency
- Implemented hybrid (dense + keyword) RAG over a vector database for domain-specific knowledge retrieval
- Integrated LLM voice commands into a ROS 2 robotic control system, including visual disambiguation: when several objects match, the user picks one by voice from the robot's map of candidates

Applied Computing and Multimedia Lab (ACM Lab), NYCU — Hsinchu, Taiwan
Research Assistant — *Feb. 2025 – Jul. 2026*

- Built a people counting and attribute analysis system: a modular pipeline of object detection, multi-object tracking and attribute recognition, with lightweight multitask CNNs for gender, age group and clothing color
- Researched state-of-the-art detection and tracking methods for real-time edge deployment
- Took part in the NTIRE 2026 Mobile Real-World Image Super-Resolution challenge with the lab's team (×4 upscaling under unknown real-world degradations, with models that must run on a phone); co-author of the challenge report
- Worked on the lab's conversational avatar with ITRI and on giving a ROS 2 robot LLM voice commands

Projects

Edge LLM Digital Avatar on NVIDIA Jetson AGX Thor↗ | Demo video↗ — *Jul. 2025 – Jul. 2026*

Technologies: Python, PyTorch, vLLM, ModelOpt (NVFP4), Docker, FastAPI, ChromaDB, faster-whisper, Kokoro TTS, MuseTalk, WebRTC, ROS 2

- Served a 26B-parameter mixture-of-experts LLM (Gemma4-26B-A4B, NVFP4) with vLLM on a single Jetson AGX Thor at ∼38 tokens/s decode, with ∼4B parameters active per token
- Fit the LLM, embedding, ASR, VLM and TTS models together in 128 GB of unified memory; benchmarked Gemma4-26B-A4B against Qwen3-30B-A3B (NVFP4)

- Kept the browser-based avatar real time: cut the time to the first spoken word to ∼1.5 s by streaming the LLM output sentence by sentence and overlapping generation with TTS; profiled per-stage latency (ASR, first LLM sentence, first TTS fragment)
- Extended the open-source TalkingHead engine with Mandarin lip-sync: server-side hanzi → pinyin → viseme generation with per-character timing, and numbers converted to hanzi so they are spoken and lip-synced
- Built the first generation as a server-rendered lip-sync video avatar (MuseTalk/Wav2Lip, emotional TTS, 25 FPS) with a FastAPI backend and multi-session WebRTC signalling; moving the face to a 3D avatar rendered in the browser freed the GPU for the LLM and made the single-board version possible

Technical Skills

Programming Languages: Python, C++, C, Java, JavaScript, SQL
Backend and Systems: FastAPI, REST APIs, WebSockets, WebRTC, multithreading, Docker, Linux, Git, ROS 2
ML Frameworks: PyTorch, Hugging Face Transformers, vLLM, OpenCV, scikit-learn
Efficient Inference: NVFP4 quantization (ModelOpt), MoE serving, visual token pruning, latency profiling, NVIDIA Jetson
AI Domains: LLMs, VLMs, RAG and vector databases, information retrieval, ASR, TTS
Spoken Languages: Spanish (native), English (advanced, TOEIC 955), Chinese (TOCFL level 3, B1)

Awards and Honors

Ranked 1st out of 166 students, NYCU Computer Science, semester 114-1 — 2025
NTU Outstanding International Graduate Student Scholarship (Master's) — 2026
MOFA Taiwan Scholarship Outstanding Academic Achievement Award — 2024, 2025
ITRI Blue Ocean Innovation Award, Innovation Challenge — 2025

Publications

J. Li, Z. Chen, K. Liu et al. (incl. **J. Tyrakowski**), "The First Challenge on Mobile Real-World Image Super-Resolution at NTIRE 2026: Benchmark Results and Method Overview"↗, *IEEE/CVF CVPR Workshops (NTIRE)*, 2026.