Hung-Yueh Chiang (江泓樂)

Mobile: +1512.825.9352 Email : hungyueh.chiang@gmail.com
Webpage | Github | LinkedIn | Google Scholar

Professional Summary

A machine learning researcher with hands-on expertise in developing **CUDA kernels** and advanced **quantization** and **compression** techniques for large language models, achieving **faster inference** and **reduced memory** usage. Skilled in deploying *optimized* **Transformers** and **State Space Models** across **edge** and **cloud** platforms, with proven gains in *speed* and *efficiency*. Passionate about transforming research innovations into scalable, production-ready AI systems.

Education

The University of Texas at Austin (UT) Aug. 2021 - Feb. 2026
Ph.D. in electrical and computer engineering
Affiliation: Energy-Aware Computing Group (EnyAC)
Research Direction: Efficient fine-tuning, model quantization, and computer vision
Thesis title: Efficient Model Adaptation and Compression for Edge Intelligence
Advisor: Prof. Diana Marculescu

National Taiwan University (NTU) Sep. 2016 - Sep. 2018
M.S. in computer science (GPA: 3.87/4.3)
Affiliation: NVIDIA-NTU AI Lab
Research Direction: 3D vision and computer vision
Thesis title: A Unified Point-Based Framework for 3D Segmentation
Advisor: Prof. Winston Hsu

ETH Zurich Jan. 2015 - Sep. 2015
Undergraduate exchange student (2 nominees in NYCU CS college)

National Yang Ming Chiao Tung University (NYCU) Sep. 2011 - Sep. 2015
B.S. in computer science (GPA: 4.08/4.3, rank 2/32)
Program of computer and electrical engineering

Industrial Experience

Senior Machine Learning Engineer, Nvidia, Santa Clara, USA Apr. 2026 – Current
- Develop W4A4 NVFP4 post-training quantization, quantization-aware training, quantization-aware distillation for LLMs
- Maintain Model-Optimizer repository

Research Scientist Intern, Eigen AI, Palo Alto CA (remote), USA Dec. 2025 – Feb. 2026
- Develop W4A4 NVFP4 post-training quantization, quantization-aware training, quantization-aware distillation for LLMs on Megatron-LM
- Deploy large-scale LLM inference with SGLang and TensorRT-LLM

Software Engineering Intern, Rivian, Palo Alto CA, USA Jun. 2023 – Aug. 2023
- Neural Architecture Search (NAS) for 3D object detection

Research Scientist Intern, Amazon, Seattle (remote), USA May 2022 – Nov. 2022
- Image synthesis and generation for shoe virtual try-on with diffusion models

Deep Learning Engineer, XYZ Robotics, Shanghai, China Jun. 2019 - May 2021

- Develop production-level deep learning vision systems on logistic robots
- Develop a multi-modal segmentation model for predicting picking areas on the objects
- Synthesize training data with Blender for unseen items to improve the model's generalization

Open-source Contributions

- TensorRT-LLM: An open-sourced library for optimizing Large Language Model (LLM) inference
- Megatron-LM: GPU-optimized library for training transformer models at scale
- Model-Optimizer: A library comprising state-of-the-art model optimization techniques
- Fast-hadamard-transform: An efficient Hadamard transform implementation
- HolisticTraceAnalysis: A PyTorch profiling tool by Facebook Research
- Elana: A Simple Energy & Latency Analyzer for LLMs

Programming Skills

- Programming languages: Python, C/C++, CUDA
- LLM serving frameworks: vLLM, TensorRT-LLM, SGLang
- LLM frameworks: Megatron-LM, Model-Optimizer, LM-Eval, ModelScope, EvalScope
- Deep learning frameworks: Pytorch, Tensorflow, MXNet, ONNX
- CUDA libraries: CUTLASS, cuBLAS, cuSPARSE, PTX
- Hardware platforms: Nvidia Jetson Series, Google Edge TPU, Intel Neural Compute Stick
- Vision/Robotic libraries: Robot Operating System, Point Cloud Library, OpenCV
- Development tools: Docker, Cmake, PyLint, Pytest, MyPy, Google Test, Git
- Web language: HTML, JQuery, Java Script, CSS
- Web framework: Django, Bootstrap, React
- 3D rendering Tools: Blender

Invited Talks

- "*Efficient Model Adaptation and Compression for Edge Intelligence*," Introduction to Machine Learning course (2025), National Tsing Hua University. Hosted by Prof. Po-An Wang.
- "*Efficient Model Adaptation and Compression*," Edge AI course (2025), National Yang Ming Chiao Tung University, Taiwan. Hosted by Prof. Kai-Chiang Wu.
- "*AI Customization: From cloud to edge - On-device transfer learning*," Edge AI course (2024), National Yang Ming Chiao Tung University, Taiwan. Hosted by Prof. Kai-Chiang Wu.
- "*AI Customization: From cloud to edge - On-device transfer learning*," Invited tutorial (2022), CyCraft Technology, Taiwan. Hosted by Dr. Chong-Kuan Chen.

Honors and Awards

- Engineering fellowship from The University of Texas at Austin graduate school, 2021
- Second place at ScanNet benchmark competition and invited talk at ScanNet Indoor Scene Understanding Challenge workshop in CVPR 2019
- Second place at SHREC17 RGB-D to CAD retrieval competition, 2017
- Taiwan Ministry of Education exchange scholarship, 2014
- Pan Wen-Yuan Foundation undergraduate scholarship (3 nominees in NCTU EE/CS), 2014
- Academic achievement award (for students at the top 5% in the class), 2014
- Research creativity award from the National Science Council, Taiwan, 2014

Patents and Disclosures

[P1] "*UniQL: Unified Quantization and Low-rank Compression for Adaptive Edge LLMs*," (**Chiang, H. Y.,** Chang, C. C., Lu, Y. C., Lin, C. Y., Wu, K. C., Abdelfattah, M. S., & Marculescu, D.), Disclosure UT Tech ID #8863 MAR, submitted Oct. 03, 2025.

[P2] "*Post-Training Quantization Recipe for Selective State Space Models*," (**Chiang, H. Y.,** Chang, C. C., Frumkin, N., Wu, K. C., Abdelfattah, M. S., & Marculescu, D.), United States Provisional Patent Application No. 63/778,277, filed March 26, 2025.

[P3] "*Quamba2: A Robust and Scalable Post-training Quantization Framework for Selective State Space Models*," (**Chiang, H. Y.,** Chang, C. C., Frumkin, N., Wu, K. C., Abdelfattah, M. S., & Marculescu, D.), Disclosure UT Tech ID #8657 MAR, submitted Feb. 19, 2025.

[P4] "*Quamba: A Post-Training Quantization Recipe for Selective State Space Models*," (**Chiang, H. Y.*,** *Chang, C. C.**, Frumkin, N., Wu, K. C., & Marculescu, D.), Disclosure UT Tech ID #8463 MAR, submitted July 11, 2024; United States Provisional Patent Application No. 63/712,949, filed October 28, 2024.

[P5] "*Efficient Low-Rank Backpropagation for Vision Transformer Adaptation*," (Yang, Y., **Chiang, H. Y.,** Li, G., Marculescu, D., & Marculescu, R.), Disclosure UT Tech ID #8355 MAR, submitted August 2024; United States Provisional Patent Application filed October 2024.

Full Publications (* Equal contribution)

Peer-Reviewed Conference Papers

[C1] **Chiang, H. Y.,** Chang, C. C., Lu, Y. C., Lin, C. Y., Wu, K. C., Abdelfattah, M. S., & Marculescu, D. (2026). *UniQL: Unified Quantization and Low-rank Compression for Adaptive Edge LLMs.* In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil.

[C2] Lu, Y. C., Yu, S. F., Weng, H. H., Wang, P. S., Hu, Y. F., Liang, H. C., **Chiang, H. Y.**, & Wu, K. C. (2026). *SkipCat: Rank-Maximized Low-Rank Compression of Large Language Models via Shared Projection and Block Skipping*. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Singapore.

[C3] **Chiang, H. Y.,** Chang, C. C., Frumkin, N., Wu, K. C., Abdelfattah, M. S., & Marculescu, D. (2025). *Quamba2: A Robust and Scalable Post-training Quantization Framework for Selective State Space Models.* In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, Vancouver, British Columbia, Canada.

[C4] **Chiang, H. Y.*,** *Chang, C. C.**, Frumkin, N., Wu, K. C., & Marculescu, D. (2025). *Quamba: A Post-Training Quantization Recipe for Selective State Space Models*. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, Singapore.

[C5] Yang, Y., **Chiang, H. Y.,** Li, G., Marculescu, D., & Marculescu, R. (2024). *Efficient Low-Rank Backpropagation for Vision Transformer Adaptation*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 36, New Orleans, Louisiana, USA.

[C6] **Chiang, H. Y.,** Frumkin, N., Liang, F., & Marculescu, D. (2023). *MobileTL: On-Device Transfer Learning with Inverted Residual Blocks*. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 37(6), pp. 7166–7174, Washington, DC, USA. (**Oral**)

[C7] **Chiang, H. Y.,** Lin, Y. L., Liu, Y. C., & Hsu, W. H. (2019). *A Unified Point-Based Framework for 3D Segmentation*. In *Proceedings of the 2019 International Conference on 3D Vision (3DV)*, pp. 155–163, IEEE, Québec City, Québec, Canada.

[C8] Lee, T., Lin, Y. L., **Chiang, H. Y.,** Chiu, M. W., Hsu, W., & Huang, P. (2018). *Cross-Domain Image-Based 3D Shape Retrieval by View Sequence Learning*. In *Proceedings of the 2018 International Conference on 3D Vision (3DV)*, pp. 258–266, IEEE, Verona, Italy. (**Oral**)

[C9] Lin, W. H., Chen, K. T., **Chiang, H. Y.,** & Hsu, W. (2018). *Netizen-Style Commenting on Fashion Photos: Dataset and Diversity Measures*. In *Companion Proceedings of The Web Conference 2018 (WWW)*, pp. 395–402, Lyon, France.

Peer-Reviewed Workshop Papers with Published Proceedings

[W1] Menn, D., Liang, F., **Chiang, H. Y.,** & Marculescu, D. (2025). *Similarity Trajectories: Linking Sampling Process to Artifacts in Diffusion-Generated Images.* In *Proceedings of the Winter Conference on Applications of Computer Vision Workshop (WACVW) on Image/Video/Audio Quality in Computer Vision and Generative AI*, Tucson, Arizona, USA.

[W2] Yang, Y., **Chiang, H. Y.,** Li, G., Marculescu, D., & Marculescu, R. (2024). *Cache and Reuse: Rethinking the Efficiency of On-Device Transfer Learning.* In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) on Efficient Deep Learning for Computer Vision*, pp. 8040–8049, Seattle, Washington, USA.

[W3] Liu, C. H., Han, Y. S., Sung, Y. Y., Lee, Y., **Chiang, H. Y.,** & Wu, K. C. (2021). *FOX-NAS: Fast, On-Device and Explainable Neural Architecture Search*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) on Low-Power Computer Vision*, pp. 789–797, Virtual.

[W4] Hua, B. S., Truong, Q. T., Tran, M. K., Pham, Q. H., Kanezaki, A., Lee, T., **Chiang, H. Y.,** … & Yeung, S. K. (2017). *SHREC'17: RGB-D to CAD Retrieval with ObjectNN Dataset*. In *Proceedings of the Eurographics Workshop on 3D Object Retrieval*, pp. 25–32, Lyon, France.

Peer-Reviewed Workshop Papers (No Published Proceedings)

[W5] Chi, T. Y., **Chiang, H. Y.,** Chang, C. C., Huang, N. C., Wu, K. C., & Marculescu, D. (2024). *QuaterMap: Efficient Post-Training Activation Pruning for Visual State Space Models*. In *3rd Workshop on Efficient Systems for Foundation Models in Forty-Second International Conference on Machine Learning (ICMLW)*, Vancouver, British Columbia, Canada.

[W6] Chi, T. Y., **Chiang, H. Y.,** Chang, C. C., Huang, N. C., & Wu, K. C. (2024). *V"Mean"ba: Visual State Space Models Only Need 1 Hidden Dimension*. In *Workshop on Machine Learning for Systems at Advances in Neural Information Processing Systems (NeurIPSW)*, Vancouver, British Columbia, Canada.

[W7] **Chiang, H. Y.,** & Marculescu, D. (2024). *SCAN-Edge: Finding MobileNet-Speed Hybrid Networks for Commodity Edge Devices.* In *5th Workshop on Practical Machine Learning for Limited/Low Resource Settings at International Conference on Learning Representations (ICLRW)*, Vienna, Austria.

[W8] Lee, K. Y., Huang, H. F., **Chiang, H. Y.,** Lee, H. C., Hsu, W. H., & Chen, W. C. (2019). *Metadata-Augmented Neural Networks for Cross-Location Solar Irradiation Prediction from Satellite Images*.

In *5th Workshop on Mining and Learning from Time Series at the Conference on Knowledge Discovery and Data Mining (KDDW)*, Anchorage, Alaska, USA.

Technical Reports (No Published Proceedings)

[T1] **Chiang, H. Y.,** Wang, B., & Marculescu, D. (2025). *ELANA: A Simple Energy and Latency Analyzer for LLMs*. arXiv preprint arXiv:2512.09946.

[T2] Liu, Y. C., Huang, Y. K., **Chiang, H. Y.,** Su, H. T., Liu, Z. Y., Chen, C. T., Tseng, C. Y., & Hsu, W. H. (2021). *Learning from 2D: Contrastive Pixel-to-Point Knowledge Transfer for 3D Pretraining*. arXiv preprint arXiv:2104.04687. (**123 citations**)

References

- **Diana Marculescu**, Professor and Chair, Department of Electrical and Computer Engineering at The University of Texas at Austin
- **Lizy Kurian John**, Professor, Department of Electrical and Computer Engineering at The University of Texas at Austin
- **Winston Hsu**, Professor, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
- **Kai-Chiang Wu**, Professor, Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
- **Peter Kuan-Ting Yu**, Chief Technology Officer (CTO) at XYZ Robotics, Shanghai, China

Students Mentored

- Chi-Chih Chiang: Ph.D. student at Cornell University
- Chi-Tien Yu: Research assistant at National Yang Ming Chiao Tung University
- Yu-Chen Lu: Ph.D. student at National Yang Ming Chiao Tung University

Academic Service

- Conference Reviewer: NeurIPS (2024, 2025), ICLR (2025, 2026), ICML (2025, 2026), AAAI (2026)
- Pro bono office hours: link