Janghwan Lee

Integrated Ph.D. Candidate
Department of Electronic Engineering
Hanyang University, Seoul 04763, Republic of Korea

hwanii0288@hanyang.ac.kr
+82-10-3220-0288
superdocker.github.io
github.com/superdocker

Research Interests

Efficient inference and training of large neural networks; post-training quantization and quantization-aware training; reduced-precision numerical formats (fixed-point, floating-point, microscaling); large language model optimization; hardware-aware algorithm design for AI accelerators.

Education

Hanyang University — Seoul, Republic of Korea
Integrated Ph.D., Department of Electronic Engineering — Mar. 2020 – Present
Advisor: Prof. Jungwook Choi (AI Hardware & Algorithm Lab)

Hanyang University — Seoul, Republic of Korea
B.S., Department of Electronic Engineering — Mar. 2014 – Feb. 2020
Advisor: Prof. Kiseok Chung Thesis: Fast Face Detector Using DCT Coefficients

Industry Experience

Research Intern — Jul. 2023 – Sep. 2023
Samsung Advanced Institute of Technology (SAIT), Suwon, Republic of Korea
Research area: Large-scale AI

Research Experience

Graduate Research Assistant — Mar. 2020 – Present
AI Hardware & Algorithm Lab, Hanyang University, Seoul, Republic of Korea
Advisor: Prof. Jungwook Choi

- **FP4 Reasoning-Accurate Quantization-Aware Training** *(ICML 2026, Spotlight).* Identified that FP4 quantization errors concentrate at low-entropy tokens (digits, operators), where noise inflates sampling errors that cascade through multi-step reasoning traces. Proposed *ReQAT*, an FP4 QAT framework with trace-aligned training (TAQ), selective entropy minimization (SEM), and quantization-friendly KV-cache initialization (Q-FIT)—surpassing BF16 reasoning accuracy with over 3× throughput speedup on production hardware.
- **Asymmetric Microscaling Floating-Point for 4-bit LLM Inference** *(ACL Findings 2025).* Found that activation outliers under group-wise 4-bit quantization introduce data asymmetry that bounds achievable accuracy, despite outlier mitigation. Proposed *AMXFP4*, which uses asymmetric shared scales for direct 4-bit casting without data rotation or calibration, outperforming MXFP4 across multi-turn conversation, long-context reasoning, and visual QA tasks.
- **Rank-Insensitive LoRA Compensation for 2-bit LLMs** *(AAAI 2025).* Identified that quantization error in 2-bit LLMs is inherently high-rank, explaining the failure of low-rank LoRA adapters for weight-level error compensation. Proposed *RILQ*, which exploits the rank-insensitive nature of model-wise activation discrepancy loss to cooperatively adjust adapters across layers, consistently improving 2-bit accuracy on LLaMA-2 and LLaMA-3 across diverse quantizers.
- **QAT via Direct Preference Optimization for Conversational LLMs** *(ACL 2024, Oral).*

Identified token-flipping caused by PTQ as the primary mechanism degrading conversational quality in LLM chatbots. Proposed *QDPO*, a novel QAT paradigm applying DPO to align quantized models with their full-precision counterparts—treating full-precision outputs as preferred responses—outperforming PTQ and knowledge-distillation baselines on two instruction-tuned LLMs in Korean and English.

- **W4A8 Quantization with Denormal Integer Format** *(EMNLP 2023).* Proposed AQAS and SLAC to enhance W4A8 PTQ by jointly handling weight-activation quantization effects and aligning calibration to target sequence lengths. Introduced *dINT*, a hybrid integer-denormal data format mitigating 4-bit weight underflow, with compatible arithmetic units achieving 2× hardware efficiency over 8-bit integer MACs on OPT and LLaMA models.
- **Mixed-Format Post-Training Quantization for Vision Transformers** *(ICASSP 2023).* Proposed an analytical framework for mixed-format (fixed/floating-point) PTQ of ViTs that selects the optimal numerical format per matrix multiplication via a simple statistical test, achieving state-of-the-art sub-8-bit accuracy across popular ViT models.
- **Sub-8-bit Floating-Point PTQ for Fine-tuned Transformers** *(AICAS 2022, Oral).* Introduced SQNR-based progressive exponent bias tuning for post-training floating-point quantization of fine-tuned transformers, achieving near full-precision BERT accuracy at 6–8 bits across GLUE and SQuAD benchmarks.
- **Reduced-Precision Simulation Frameworks.** Led development of *QLLM-INFER* (2025) in collaboration with Dnotitia Inc.—an open-source LLM inference simulation framework evaluating 8 quantization methods on LLaMA-3.1-8B across weight-activation, weight-only, and KV-cache quantization schemes (`github.com/dnotitia/qllm-infer`). Developed *MX-QLLM* (2024)—an open-source reduced-precision MX-format LLM inference framework supporting FP4/FP6/FP8 element types with configurable shared-scale strategies (PoT, FP8, and asymmetry) and randomized Hadamard rotation for outlier-free quantization (`github.com/aiha-lab/MX-QLLM`). Built a PyTorch CUDA-based reduced-precision training simulator (2021) with configurable precision for weights, activations, gradients, and accumulators; validated with no accuracy drop on SSD-Lite and ResNet/MobileNet at 8-bit.

Publications

⋆ *denotes equal contribution.* ***Boldface*** *indicates the author.*

[1] **Janghwan Lee**, Sihwa Lee, Jinseok Kim, Yongjik Kim, Jieun Lim, Jinwook Oh, and Jungwook Choi. "ReQAT: Achieving Full-Precision Reasoning Accuracy with 4-bit Floating-Point Quantization-Aware Training." *International Conference on Machine Learning (ICML)*, 2026. **[Spotlight]**

[2] **Janghwan Lee**, Jiwoong Park, Jinseok Kim, Yongjik Kim, Jungju Oh, Jinwook Oh, and Jungwook Choi. "AMXFP4: Taming Activation Outliers with Asymmetric Microscaling Floating-Point for 4-bit LLM Inference." *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025.

[3] Geonho Lee⋆, **Janghwan Lee**⋆, Sukjin Hong⋆, Minsoo Kim, Euijai Ahn, Du-Seong Chang, and Jungwook Choi. "RILQ: Rank-Insensitive LoRA-based Quantization Error Compensation for Boosting 2-bit Large Language Model Accuracy." *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[4] **Janghwan Lee**⋆, Seongmin Park⋆, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. "Improving Conversational Abilities of Quantized Large Language Models via Direct Preference Alignment." *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. **[Oral]**

[5] Dong-eon Won⋆, Yeeun Kim⋆, **Janghwan Lee**, Minjae Lee, Jonghyun Bae, Jongjoo Park, Jeongyong Song, and Jungwook Choi. "ISP2DLA: Automated Deep Learning Accelera-

tor Design for On-Sensor Image Signal Processing." *IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP)*, 2024.

[6] Minjae Lee, Seongmin Park, Hyungmin Kim, Minyong Yoon, **Janghwan Lee**, Junwon Choi, Nam Sung Kim, Mingu Kang, and Jungwook Choi. "SPADE: Sparse Pillar-based 3D Object Detection Accelerator for Autonomous Driving." *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024.

[7] Youngdeok Hwang*, **Janghwan Lee***, Jiwoong Park, Jieun Lim, and Jungwook Choi. "Searching Optimal Floating-Point Format for Sub-8-Bit Large Language Model Inference." *International Conference on Electronics, Information, and Communication (ICEIC)*, 2024. **[Oral]**

[8] **Janghwan Lee***, Minsoo Kim*, Seungcheol Baek, Seok Joong Hwang, Wonyong Sung, and Jungwook Choi. "Enhancing Computation Efficiency in Large Language Models through Weight and Activation Quantization." *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[9] Minsoo Kim, Sihwa Lee, **Janghwan Lee**, Sukjin Hong, Du-Seong Chang, and Wonyong Sung, and Jungwook Choi. "Token-Scaled Logit Distillation for Ternary Weight Generative Language Models." *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[10] **Janghwan Lee**, Youngdeok Hwang, and Jungwook Choi. "Finding Optimal Numerical Format for Sub-8-bit Post-Training Quantization of Vision Transformers." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[11] Janghyeon Kim, **Janghwan Lee**, JeongHo Han, Sangheon Lee, and Jungwook Choi. "Range-Invariant Approximation of Non-Linear Operations for Efficiently Fine-tuning BERT." *ACM/IEEE Design Automation Conference (DAC)*, 2023.

[12] **Janghwan Lee** and Jungwook Choi. "Optimizing Exponent Bias for Sub-8bit Floating-Point Inference of Fine-tuned Transformers." *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2022. **[Oral]**

Honors & Awards

- **Excellence in Research and Education Award**, ENRICH IT Award, 4th Stage BK21 2025
- **First Place**, Model Compression Track, AI Grand Challenge, Ministry of Science and ICT, Korea 2020
- **Integrated Ph.D. Course Scholarship** (Full Tuition), Hanyang University 2020–2022
- **Research Scholarship** (KRW 24M total), ISRC 2020–2022
- **B.S. Course Scholarship** (Full Tuition), Hanyang University 2014–2015, 2018–2019

Technical Skills

Languages: Python, C, C++

Frameworks: PyTorch, Hugging Face Transformers

Platforms: Linux, Git, Docker