

Minsoo Kim

+82-10-8203-6871 | minsoo2333@hanyang.ac.kr | marsjacobs.github.io |  Google Scholar

RESEARCH INTERESTS

Efficiency for LLM Inference - long context optimization for LLMs/MLLMs; long video understanding; model quantization; knowledge distillation; parameter efficient fine-tuning

EDUCATION

- **Hanyang University** Mar. 2021 - Feb. 2026 (expected)
Ph.D. Student in Electronic Engineering Seoul, South Korea
 - Advisor: Professor [Jungwook Choi](#)
 - [Artificial Intelligence Hardware & Algorithm Lab](#)
- **Hanyang University** Feb. 2021
B.S. in Electronic Engineering Seoul, South Korea
 - Thesis: Improving training method for very low bit weight quantization of Light Deep Learning Model
 - Advisor: Professor [Jungwook Choi](#)

EXPERIENCE

- **Apple AIML** Mar. 2025 - Sep. 2025
DMLI Research Intern Seattle, US
- **Qualcomm AI Research** Mar. 2024 - Mar. 2025
Research Intern Seoul, Korea

PUBLICATIONS

C=CONFERENCE, S=IN SUBMISSION

- [S.1] **Minsoo Kim**, Arany Kundu, Han-Byul Kim, Richa Dixit, and Minsik Cho. [EpiCache: Episodic KV Cache Management for Long Conversational Question Answering](#). Preprint, 2025.
- [C.1] **Minsoo Kim**, Kyuhong Shim, Jungwook Choi, and Simyung Chang. [InfiniPot-V: Memory-Constrained KV Cache Compression for Streaming Video Understanding](#). *Thirty-ninth Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [C.2] Seunghan Yang, Juntae Lee, Jihwan Bang, Kyuhong Shim, **Minsoo Kim**, and Simyung Chang. [Learning Contextual Retrieval for Robust Conversational Search](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.
- [C.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](#). In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2025.
- [C.4] **Minsoo Kim**, Kyuhong Shim, Jungwook Choi, and Simyung Chang. [InfiniPot: Infinite Context Processing on Memory-Constrained LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.
- [C.5] **Minsoo Kim**, Sihwa Lee, Wonyong Sung and Jungwook Choi. [RA-LoRA: Rank-Adaptive Parameter-Efficient Fine-Tuning for Accurate 2-bit Quantized Large Language Models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*.
- [C.6] 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](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [C.7] **Minsoo Kim**, Sihwa Lee, Jangwhan Lee, Sukjin Hong, Du-Seong Chang, Wonyong Sung and Jungwook Choi. [Token-Scaled Logit Distillation for Ternary Weight Generative Language Models](#). In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [C.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](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. (*Co-First author)
- [C.9] **Minsoo Kim**, Kyuhong Shim, Seongmin Park, Wonyong Sung and Jungwook Choi. [Teacher Intervention: Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

- [C.10] **Minsoo Kim**, Sihwa Lee, Sukjin Hong, Du-Seong Chang, and Jungwook Choi. **Understanding and Improving Knowledge Distillation for Quantization-Aware Training of Large Transformer Encoders**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- [C.11] Joonsang Yu, Junki Park, Seongmin Park, **Minsoo Kim**, Sihwa Lee, Donghyun Lee, Jungwook Choi. **NN-LUT: neural approximation of non-linear operations for efficient transformer inference**. In *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, 2022.

RESEARCH EXPERIENCE

Research Intern, Apple MIND (Advisor. Minsik Cho, Richa Dixit)

- Episodic KV Cache Management for Long-Term Multi-Turn Conversation - [S.1]
 - Clustering based episodic KV cache compression with constrained memory usage
 - 40% accuracy improvement, 2.4/3.5x latency/peak memory reduction up to 100K conversation





Research Intern, Qualcomm AI Research (Advisor. Simyung Chang)

- Continual KV Cache Compression for Memory-Constrained Streaming Video Understanding - [C.1]
 - Training-free video-based KV cache compression method with spatiotemporal importance scoring
 - Achieve 94% KV cache compression while maintaining accuracy for long video understanding
- Infinite Context Distillation for Memory-Constrained LLMs - [C.4]
 - Chunk processing based KV cache control framework enabling infinite context distillation
 - Up to 8x compression with memory-constrained LLaMA/Mistral/Gemma KV cache compression

Research Assistant, Hanyang University (Advisor. Prof. Jungwook Choi)

- LLM Quantization Error Compensation with Parameter Efficient Fine Tuning (LoRA)
 - Rank-insensitive low-bit quantization error compensation with loss objective exploration - [C.3]
 - Analyze high-rank characteristics of low-bit quantization error with rank-adaptive LoRA - [C.5]
- LLM Quantization (Quantization-Aware Training - QAT, Post-Training Quantization - PTQ)
 - Probabilistic confidence-based token-scaling KD technique for LLM 2-bit (ternary) QAT - [C.7]
 - 4-bit weight and 8-bit activation PTQ based on comprehensive analysis of LLM quantization effects - [C.8]
- Transformer Encoder (BERT/Roberta/ViT) QAT with Knowledge Distillation (KD)
 - Teacher-forced KD technique in BERT and ViT for speed-up fine-tuning time up to 12.5x - [C.9]
 - Low-bit quantization effects on self-attention block in Transformer encoders over NLU tasks - [C.10]

HONORS AND AWARDS

- | | |
|--|--|
| • Outstanding Reviewer
EMNLP 2024 | November 2024
 |
| • AICAS Grand Challenge 2024
3rd place, SW&HW Co-Optimization for LLM | March 2024
 |
| • Qualcomm Innovation Fellowship Korea 2023
Winner, Qualcomm AI Research | November 2023
 |
| • AI Grand Challenge
1st place, Korea Ministry of Science and ICT | November 2020
 |

TEACHING EXPERIENCE

- | | |
|---|-------------|
| • Teaching Assistant - SOC Design
Hanyang University | Spring 2021 |
| • Teaching Assistant - Introduction to SW Optimization
Hanyang University | Fall 2023 |

OTHER EXPERIENCES

- **Academic Services:**
Reviewer for ACL Rolling Review (ARR), NeurIPS, ICLR, ICML, COLM, AAAI, TAI
2023 - present
- **Academic Volunteer:**
EMNLP Student Volunteer Program
2022 - 2024
- **English:**
Served as a KATUSA (Korean Agumentation to the US Army)
July 2017 - April 2019