

Woosang Lim

Seoul, South Korea | ftyg656512@snu.ac.kr | [linkedin.com/in/woosang-lim-14b48a229](https://www.linkedin.com/in/woosang-lim-14b48a229) | [Google Scholar](#)

RESEARCH PROFILE

Ph.D. student at Seoul National University researching efficient multi-agent systems, LLM quantization, KV-cache quantization, LLM pruning, and diffusion LLMs.

EDUCATION

Seoul National University <i>Ph.D. Student, Interdisciplinary Program in Artificial Intelligence</i>	Seoul, South Korea Mar. 2024 –
Seoul National University <i>B.S. in Electrical and Computer Engineering</i>	Seoul, South Korea Mar. 2017 – Feb. 2024

RESEARCH EXPERIENCE

Graduate Researcher <i>Seoul National University, Computing and Memory Architecture Laboratory</i> Advisor: Prof. Sungjoo Yoo	Present Seoul, South Korea
---	-------------------------------

WORK EXPERIENCE

Internship <i>SK Hynix CIS Process Integration</i> Understand 300mm ToF Sensors and their core technologies	2022 Summer Icheon, South Korea
---	------------------------------------

PUBLICATIONS

[arXiv] Unlocking Structured Pruning for Reasoning Language Models via Attention-Preserving MLP Trimming Woosang Lim*, Haneul Yoo*, Vincent-Daniel Yun*, and Sungjoo Yoo. <i>arXiv:2610.19121</i>.	2026
[arXiv] Prefill-Free Cross-Family KV Cache Transfer for Heterogeneous Multi-Agent LLMs Vincent-Daniel Yun*, Woosang Lim*, Haneul Yoo, Sungjoo Yoo, Murali Annavaram, and Sai Praneeth Karimireddy. <i>arXiv:2609.32259</i>.	2026
[arXiv] Output-Aware Rotation for INT2 KV-Cache Quantization Vincent-Daniel Yun*, Woosang Lim*, Minsoo Cheong, Sunwoo Lee, Murali Annavaram, Sai Praneeth Karimireddy, and Sungjoo Yoo. <i>arXiv:2608.02691</i>.	2026
[EMNLP 2026] Locality-Aware Redundancy Pruning for LLM Depth Compression Vincent-Daniel Yun*, Youngrae Kim*, Woosang Lim*, Youngjin Heo, Minkyu Kim, and Sunwoo Lee.	2026
[ICML 2026 AdaptFM] Rethinking Layer Redundancy in Large Language Models: Calibration Objectives and Search for Depth Pruning Minkyu Kim*, Vincent-Daniel Yun*, Youngrae Kim*, Youngjin Heo, Suin Cho, Seong-hun Kim, Woosang Lim, and Gaoul Kwon.	2026
[NeurIPS 2026] EntropyCache: Decoded Token Entropy Guided KV Caching for Diffusion Language Models Minsoo Cheong, Donghyun Son, Woosang Lim, and Sungjoo Yoo.	2026
[ACL SRW 2025] Grouped Sequency-arranged Rotation: Optimizing Rotation Transformation for Quantization for Free Euntae Choi*, Sumin Song*, Woosang Lim*, and Sungjoo Yoo. pp. 165–172.	2025
[EMNLP Findings 2025] Rotate, Clip, and Partition: Towards W2A4KV4 Quantization by Integrating Rotation and Learnable Non-uniform Quantizer Euntae Choi*, Sumin Song*, Woosang Lim, and Sungjoo Yoo. pp. 7568–7590.	2025

* denotes equal contribution.

HONORS AND AWARDS

Scholarships

- Merit-based Academic Scholarship, Seoul National University (2022-1, 2022-2)

TEACHING EXPERIENCE

Teaching Assistant

Seoul National University

- 4190.309A 001: Hardware System Design (2024 Spring)
- 4190.308 001: Computer Architecture (2024 Fall)
- 4190.571 001: Advanced Computer Architecture (2025 Spring, 2026 Spring)