Minseo Kim

ms05251@snu.ac.kr | Website | minseo25 | LinkedIn

Research Interests

AI & Machine Learning Systems, Efficiency, HW–SW Co-Design

System-aware efficiency for large-scale ML models, focusing on resource bottlenecks and HW–SW co-design for practical deployment.

Education

- **Seoul National University** — *Mar 2020 – Aug 2026 (Expected)*
  B.S. in Computer Science and Engineering, Summa cum laude (expected) — Seoul, South Korea
  - GPA: **4.19 / 4.30 (major)**, 4.11 / 4.30 (overall)
  - 2-year absence to fulfill military duty (Oct. 2021 - Jul. 2023)
- **University of California, Berkeley** — *Jun 2025 – Aug 2025*
  Visiting student — Berkeley, CA, USA
  - GPA: 4.00 / 4.00; Courses: Computer Security (A+), Introduction to Artificial Intelligence (A+)
- **Korea Minjok Leadership Academy** — *Mar 2017 – Feb 2020*
  *Secondary Education, **First class honor** in STEM field (1st / 73)* — Gangwon-do, South Korea

Publications

C=Conference, S=In Submission

* Equal contribution

[C.1] Taebaek Hwang*, Minseo Kim*, Gisang Lee, Seonuk Kim, Hyunjun Eun (2025). **KRETA: A Benchmark for Korean Reading and Reasoning in Text-Rich VQA Attuned to Diverse Visual Contexts**. In Proceedings of *EMNLP 2025*. [Paper] [Project Page][Hugging Face]

[C.2] Bonggeun Sim, Yushin Kim, Minseo Kim, Yeonhong Park, Jae W. Lee (2025). **InstANNS: Scalable Approximate Nearest Neighbor Search via Cost-Efficient In-Storage Processing**. In Proceedings of *CIKM 2025*. [Paper]

[S.1] Minseo Kim, Chenfeng Xu, Coleman Hooper, Harman Singh, Ben Athiwaratkun, Ce Zhang, Kurt Keutzer, Amir Gholami (2025). **CDLM: Consistency Diffusion Language Models for Faster Sampling**. Submitted to *MLSys 2026*. [arXiv]

Preprints

[1] Minseo Kim, Coleman Hooper, Aditya Tomar, Chenfeng Xu, Mehrdad Farajtabar, Michael W. Mahoney, Kurt Keutzer, Amir Gholami (2025). **Beyond Next-Token Prediction: A Performance Characterization of Diffusion versus Autoregressive Language Models**. *arXiv preprint*. [arXiv]

Research Experience

- **Berkeley Artificial Intelligence Research (BAIR) Lab [🌐]** — *Jun 2025 – Present*
  Undergraduate Visiting Researcher (Advisor: Prof. Kurt Keutzer) — Berkeley, CA, USA
  - LLM inference on compute-in-memory (CIM) hardware
    - Proposed and implemented **KV-cache compression** via offline dictionary learning and online sparse coding for memory-constrained CIM deployments.
    - Integrated adaptive hierarchical sparsity and query-aware KV reconstruction, reducing dynamic memory loads by 15× while maintaining performance across LongBench tasks.
  - Efficient inference in **Diffusion Language Models (DLMs)**
    - Characterized DLMs via GPU roofline analysis to evaluate implications for block-wise and batched decoding, identifying excessive refinement steps and caching incompatibility as primary bottlenecks.
    - Developed a consistency-distilled DLM with block-causal attention to enable KV caching, reducing refinement steps by 3.4×–7.9× and latency by 3.6×–14.5× while maintaining accuracy on math/code benchmarks.
- **SNU Architecture and Code Optimization Lab [🌐]** — *Jan 2025 – Feb 2025*
  Undergraduate Researcher (Advisor: Prof. Jae W. Lee) — Seoul, South Korea
  - Integrated SPDK into the existing ANNS system and optimized it, resulting in a 2.15× increase in throughput.
  - Built the host ↔ SSD interface for a new NVMe command enabling in-storage PQ filtering, cutting host I/O.

Work Experience

- **Deeping Source Inc. [🌐]** — *Jul 2024 – Aug 2024*
  ML Engineering Intern — Seoul, South Korea
  - Added a gaze-vector inference capability to an existing ML model for pedestrian attribute recognition, achieving a 90% soft-hit rate.
  - Generated training data using IMU sensors and built a robust data pipeline for large-scale data acquisition.

Projects

- **Open-Source Contribution, Samsung NNTrainer** *Oct 2025 – Nov 2025*
  Tools: C++
  - Enhanced NNTrainer's on-device training stack by adding gradient checkpointing and lightweight optimizers (Lion, Sophia) for improved memory and training efficiency.
- **ARC-AGI (Abstraction & Reasoning Corpus) Solver** *Mar 2025 – Jun 2025*
  Tools: PyTorch [ ]
  - Intro to Deep Learning term project: built an ARC-AGI solver; placed **1st** among 35 teams on the hidden evaluation.
  - Fine-tuned Qwen3-4B with LoRA, curated data augmentation, and a custom lm_head; two test-time scaling methods (test-time training and grid-wise voting). (Technical report: link)
- **KRETA: Korean Text-Rich VQA Benchmark & VLM Fine-Tuning** *Oct 2024 – Feb 2025*
  Tools: PyTorch [ ] [ ]
  - Collected Korean text-rich image datasets and fine-tuned LLaVA-Onevision to strengthen Korean capability.
  - Built an end-to-end generation pipeline and released a high-quality Korean text-rich VQA benchmark.
- **Art College Graduation Exhibition Sales Website** *Sep 2024 – Dec 2024*
  Tools: Django, React, Supabase, AWS [ ] [⌂]
  - Designed a Supabase-based database, implemented RESTful APIs with Django, integrated a KakaoPay module for payment processing, and deployed the service on AWS.
  - Collaborated with the Metal Craft and Ceramics departments at SNU for product registration.
- **Window Software Vulnerability Research & Exploitation** *Sep 2023 – Feb 2024*
  Tools: Windbg, IDA, Pwndbg, Fuzzer, x86-64 Assembly
  - Completed intensive training at WhiteHat School (web, system, cloud security, forensics) and researched Windows application vulnerabilities using static/dynamic analysis and fuzzing.
  - Responsibly disclosed **three CVE-registered systems vulnerabilities** via ZDI (CVE-2024-11510, CVE-2024-11511, CVE-2024-11512); coordinated with vendors to ensure patch releases.

Honors and Awards

- **National IT Industry Promotion Agency (NIPA) President's Award - 2025 AI Chip Contest** *Dec 2025*
  - Grand Prize (top award) in the AI semiconductor application track for optimizing vision pre-processing workloads on Furiosa/Rebellions NPUs (KRW 10M prize).
- **Gaheonsindo Foundation Scholarship (Full Undergraduate Tuition)** *Mar 2024 – Present*
- **SNU Specialized Semiconductor College Scholarship** *Feb 2024 – Present*
- **Accelerator Programming Winter School Outstanding Graduate** *Feb 2025*
  - Secured 1st place in CUDA-based heterogeneous and accelerator computing competitions.
- **Ministry of Science and ICT Award - Whitehat School Outstanding Graduate** *Mar 2024*
  - Ranked 1st out of 309 participants (KRW 3M prize).
- **Merit-based Scholarship** *2021, 2023*

Leadership & Teaching

- **President, Guardian (SNU CSE Security Research Club)** *Aug 2024 – Aug 2025*
  - Organized and delivered weekly seminars for new members on system hacking, reversing, web security, and cryptography; designed onboarding curriculum and mentorship.
- **Presenter (28th Winter Hacking Camp) – Accepted Talk** *Feb 2024*
  - Delivered a 1-hour instructional session on Windows software vulnerability detection and analysis to 60 participants.
- **Team WWW Lead (Whitehat School)** *Oct 2023 – Jan 2024*
  - Led an outstanding project team; discovered three CVE-assigned vulnerabilities and concluded with a poster presentation.

Skills

- **Programming Languages:** Python, x86-64 Assembly, C, C++, Java
- **AI/ML:** PyTorch, CUDA, Triton, OpenCV, ClearML
- **DevOps & Version Control:** Git, Docker
- **Other Tools & Technologies:** IDA, WinDbg, Pwndbg, Fuzzer, Unity, React, Django, Supabase, MySQL

Language Proficiency

Korean (Native), English (Fluent)