



Faculty



송현오 교수님

- Associate Professor, SNU
- Research Scientist, Google Research
- Postdoc, Stanford University
- Ph.D. in Computer Science, UC Berkeley

Students



6 MS/Ph.D. students

- 3 B.S. in Computer Science and Engineering
- 1 B.S. in Mathematical Science
- 1 B.S. in Statistics
- 1 B.S. in Economics

Alumni (Ph.D.)

- Deokjae Lee Ph.D. (2020 ~ 2026)
- Jang-Hyun Kim Ph.D. (2019 ~ 2025)
- Now at Apple
- Yeonwoo Jeong Ph.D. (2018 ~ 2024)
- Now at Ludo Robotics
- Jaekyeom Kim Ph.D. (2018 ~ 2023)
- Now at LG AI Research

Alumni (M.S.)

- Wonho Choo (Kakao), Hyungseok Kim (NVIDIA), Junhyeok Lee, Wun Hur

Publication

We publish papers on major machine learning conferences

- In 2026, 2 ICML, 1 EMNLP
- In 2025, 1 ICML, 1 TMLR, 3 NeurIPS (1 Oral)
- In 2024, 1 ICLR, 2 ICML
- In 2023, 1 ICML, 1 ACL, 3 NeurIPS
- In 2022, 2 ICML, 1 NeurIPS, 1 AAAI, 1 AISTATS
- In 2021, 1 ICLR (oral), 1 NeurIPS
- In 2020, 1 ICML
- In 2019, 3 ICML (2 long talk), 1 CVPR
- In 2018, 1 ICML (1 long talk)

Environment

Lab office

- Samsung Electronics-SNU Research Center (7F)
- Haedong Advanced Engineering Center (6F, Building 303)



Research

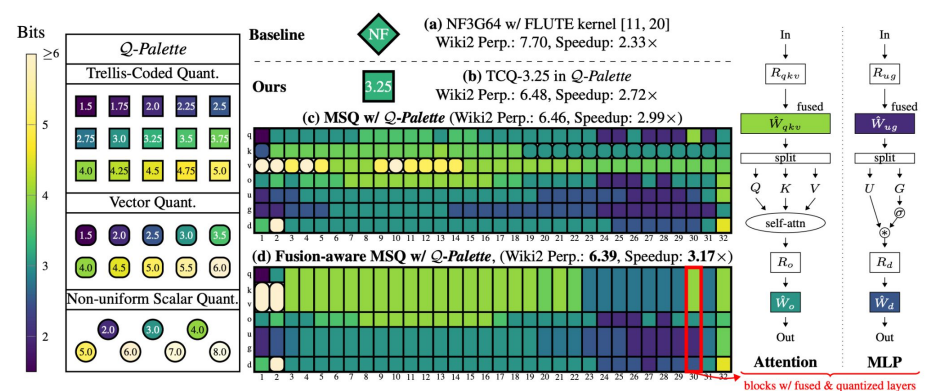
LLM Efficiency

- KV Cache Optimization [NeurIPS 2025 **Oral** (77/21,575=0.35%)]

Context: ExLlamaV3 GitHub Repo (480K tokens)		
Model: Qwen2.5-72B-Instruct-1M (Bfloat16, released earlier than ExLlamaV3)		
Q ₁ : What must max_num_tokens be a multiple of when creating a cache?	(A: 256)	
Q ₂ : What bit ranges are allowed for keys and values in quantized cache layers?	(A: From 2 to 8 bits)	
Q ₃ : Which C++/CUDA file handles the implementation of dequant_cache_paged?	(A: exllamav3/exllamav3_ext/cache/q_cache.cu)	
Only README.md	Full KV Cache	KVzip
✗ A ₁ : a multiple of 4096	✓ A ₁ : 256	✓ A ₁ : 256
✗ A ₂ : 4bpw	✓ A ₂ : Keys: 2..8 bits Values: 2..8 bits	✓ A ₂ : Keys: 2..8 bits Values: 2..8 bits
✗ A ₃ : dequant_cache_paged is implemented in src/dequant.cpp	✓ A ₃ : exllamav3/exllamav3_ext/cache/q_cache.cu	✓ A ₃ : exllamav3/exllamav3_ext/cache/q_cache.cu
Memory (KV): 0.1 GB Speed (decoding): 8.4 ms/token	Memory (KV): 27.5 GB Speed (decoding): 49.0 ms/token	Memory (KV): 8.2 GB Speed (decoding): 24.1 ms/token

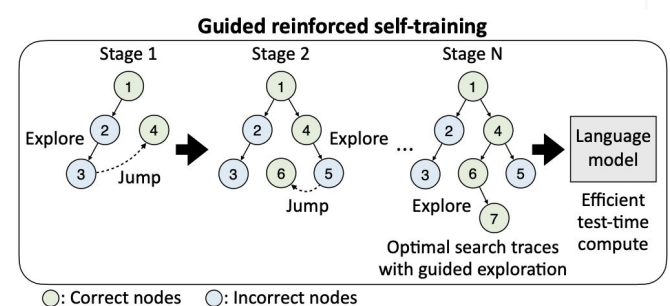
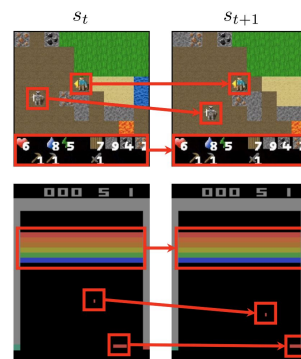
- Mixed Precision Quantization [NeurIPS 2025]

$$\begin{aligned} & \underset{P_{gq} \in \{0,1\}}{\text{minimize}} && \sum_{g \in \mathcal{G}} \sum_{q=1}^{|\mathcal{Q}|} P_{gq} \cdot \sum_{l \in \mathcal{g}} \ell_{lq} \\ & \text{subject to} && \sum_{g \in \mathcal{G}: l \in \mathcal{g}} \sum_{q=1}^{|\mathcal{Q}|} P_{gq} = 1, \quad \forall l \in \bigcup_{b=1}^B \{q_b, k_b, v_b, o_b, u_b, g_b, d_b\}, \quad \sum_{g \in \mathcal{G}} \sum_{q=1}^{|\mathcal{Q}|} P_{gq} \cdot c_{gq} \leq C, \end{aligned}$$



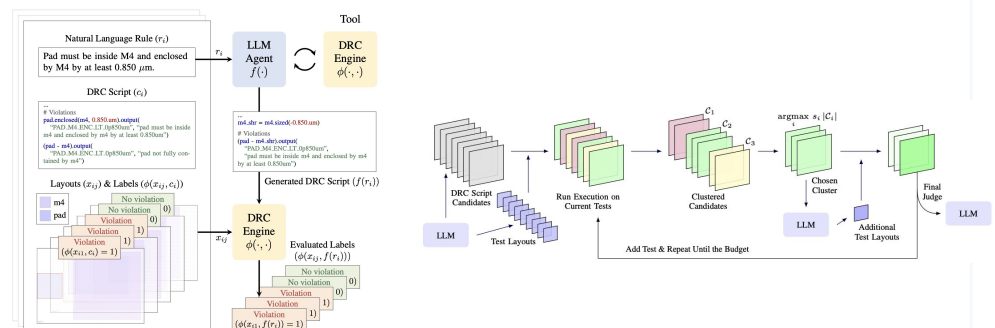
Reinforcement Learning

- World Model [ICML 2026], LLM Reasoning [NeurIPS 2025]



LLM Agents

- Coding with domain-specific language [ICML 2026]



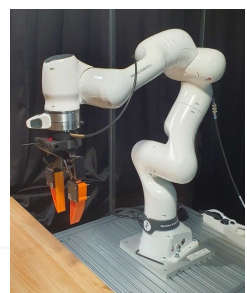
Resources

Robot

- Franka Research 3

Lab server

- 50+ servers with 200+ GPUs
- Slurm job manager



2026.09