Yoonjeon **Kim**

PhD Student

My research focuses on reinforcement learning (RL)-based post-training methods to enhance reasoning robustness in both language and multi-modal models. Specifically, I am exploring RL fine-tuning techniques for large language models (LLMs) aimed at efficient inference and evaluation at rollout phase, and also extending the reasoning ability of (M)LLMs within RAG systems.

- **Efficient RL-based Post-training**: Efficient Inferences in RL training for reasoning model.
- **Multi-modal Reasoning**: Visual question answering enhancement by learning from similar problem solving principles in RAG.

Contact

Seoul, South Korea
+82 1026199803
yoonkim313@kaist.ac.kr
akatigre
yoonjeon-kim-9898061b0
0000-0001-6545-5274

Skills

Programming

Python | Java | R

AI Frameworks

Ray | vLLM | VeRL

Languages

English	(Fluent)
Korean	(Native)
Spanish	(Basic)

Education

KAIST — *Daejeon, South Korea*
PhD (Candidate) — *Mar.2023 - Now*
- Major in Artificial Intelligence.
- Advisor: Eunho Yang

KAIST — *Daejeon, South Korea*
Master — *Mar.2021 - Feb.2023*
- Major in Artificial Intelligence.
- Advisor: Eunho Yang

Yonsei University — *Seoul, South Korea*
Bachelor — *Mar.2017 - Feb.2021*
- Major in Applied Statistics.
- Full Scholarship

Invited Talks

AI EXPO KOREA KAIST AI Tech Seminar — *KOEX, Seoul*
Presentor — *2025*

Experience

Internship, Machine Learning Researcher — *Bundang, South Korea*
NAVER Cloud — *Jul 2023 - Oct 2023*
- Main achievements:
  - 2D Casual Video to 3D Dynamic Neural Radiance Fields.
  - 3D Reconstrucion from Monocular Video with Generation Model.
  - Pretrained Diffusion based NeRF Model Distillation.

Publications

Reasoning Model is Stubborn: Diagnosing Instruction Overriding in Reasoning Models *Arxiv 2025*

Doohyuk Jang*, **Yoonjeon Kim***, Chanjae Park, Hyun Ryu, Eunho Yang (*Equal Contribution)

- [paper] [Code] [Project Page]

Preserve or Modify? Context-Aware Evaluation for Balancing Preservation and Modification in Text-Guided Image Editing *CVPR 2025*

Yoonjeon Kim*, Soohyun Ryu*, Yeonsung Jung, Hyunkoo Lee,
Joowon Kim, June Yong Yang, Jaeryong Hwang, Eunho Yang (*Equal Contribution)

- [paper] [Code] [Project Page]

Learning Input-Agnostic Manipulation Directions in StyleGAN with Text Guidance *ICLR 2023*

Yoonjeon Kim, Hyunsu Kim, Junho Kim, Yunjey Choi ,Eunho Yang

- [paper] [Code]

Sequential targeting: A continual learning approach for data imbalance in text classification *ESWA 2021*

Joel Jang, **Yoonjeon Kim**, Kyoungho Choi, Sungho Su

- Journal of Expert System with Applications, Oct 2021.
- [Paper]

Project

A Study on Optimization and Network Interpretation Method for Large-Scale Machine Learning *Mar.2024 - Feb.2027*

National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)

- Investigating optimization techniques and interpretability methods for large-scale machine learning models
- Focus on scalable algorithms and transparency in deep neural networks

Efficient Foundation Models on Intel Systems *Sep 2024 - Aug 2025*

Intel Corporation & NAVER

- Development of efficient and scalable foundation models optimized for Intel hardware platforms
- Research on model compression, hardware-aware training, and deployment strategies

Hyper-Scale AI Foundation Models *Sep 2021 - Aug 2023*

NAVER

- Development of hyper-scale AI models for visual generative models