

Yoonjeon Kim

Ph.D. Candidate in Artificial Intelligence, KAIST

yoonkim313@kaist.ac.kr — [Github](#) — [Google Scholar](#) — [Website](#)

Expected Graduation: Feb. 2027

Summary

Ph.D. candidate at KAIST, expected to graduate in Feb. 2027, working on RL-based post-training for reasoning robustness in language and multimodal models. First-author publications at ICML, CVPR, and ICLR, with research spanning reasoning-model self-verification, instruction-following robustness, text-guided image editing evaluation, StyleGAN manipulation, and multimodal discrete diffusion. Former research intern at Sony Research, working on reinforcement learning for multimodal discrete diffusion models.

Research Interests

- **RL-based post-training:** Self-rewarding and weakly supervised RL methods for improving reasoning robustness, self-verification, and instruction adherence in reasoning models.
- **Multimodal reasoning with discrete diffusion:** RL post-training and data curation methods for improving visual question answering and multimodal reasoning through analogous problem-solving signals.
- **Model evaluation and interpretability:** Evaluation pipelines for diagnosing hallucination, instruction overriding, preservation/modification trade-offs, and reasoning failures in multimodal and generative models.

Experience

Sony Research

Research Intern

Tokyo, Japan

Jan. 2026 – Jun. 2026

- Mentored by [Yuhta Takida](#) and [Chieh-Hsin \(Jesse\) Lai](#) on multimodal discrete diffusion and RL-based post-training for reasoning.
- Developed RL post-training methods for multimodal discrete diffusion models, targeting visual-textual reasoning and visual question answering.

NAVER Cloud

Machine Learning Research Intern

Bundang, South Korea

Jul. 2023 – Oct. 2023

- Researched 2D video to 3D dynamic Neural Radiance Fields for monocular-video 3D reconstruction.
- Investigated pretrained diffusion-based NeRF model distillation for efficient 3D scene representation and generation.

Publications

Verifying Meta-Awareness via Predictive Rewards in Reasoning Models

[\[Paper\]](#) [\[Code\]](#)

Yoonjeon Kim*, Doohyuk Jang*, Eunho Yang (*Equal contribution)

International Conference on Machine Learning (ICML), 2026

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

[\[Paper\]](#) [\[Code\]](#) [\[Project Page\]](#)

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

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

Learning Input-Agnostic Manipulation Directions in StyleGAN with Text Guidance

[\[Paper\]](#) [\[Code\]](#)

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

International Conference on Learning Representations (ICLR), 2023

Sequential Targeting: A Continual Learning Approach for Data Imbalance in Text Classification

[\[Paper\]](#)

Joel Jang, Yoonjeon Kim, Kyoungcho Choi, Sungho Suh

Expert Systems with Applications

Under Review

Efficient Reinforcement for Visual-Textual Thinking with Discrete Diffusion Model

[\[Paper\]](#)

Yoonjeon Kim, Yuhta Takida, Chieh-Hsin Lai, Eunho Yang, Yuki Mitsufuji

Under Review

Reasoning Model is Stubborn: Diagnosing Instruction Overriding in Reasoning Models [\[Paper\]](#) [\[Code\]](#) [\[Project Page\]](#)

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

Under Review

Skills

Programming	Python, R
ML Frameworks	PyTorch, Transformers, DeepSpeed, FSDP
RL & Post-training	VeRL, TRL, vLLM, Ray, GRPO, PPO, DPO, RLVR
Generative Models	Diffusion models, discrete diffusion, GANs, DiT, NeRF
Infrastructure	Multi-node GPU training, Slurm, Docker, Weights & Biases, Git, Kubernetes
Evaluation	VLM evaluation, visual-spatial reasoning, instruction-following robustness, dataset curation
Languages	Korean (native), English (fluent)

Selected Research Projects

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

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

KAIST

- Investigating optimization and interpretability methods for large-scale machine learning models, with emphasis on scalable algorithms and transparent model behavior.

Efficient Foundation Models on Intel Systems

Sep. 2024 – Aug. 2025

Intel Corporation & NAVER

KAIST / NAVER

- Researched efficient and scalable foundation models optimized for Intel hardware platforms.

Hyper-Scale AI Foundation Models

Sep. 2021 – Aug. 2023

NAVER

KAIST / NAVER

- Developed visual generative-model components for hyper-scale AI foundation models, focusing on text-guided image editing.

Education

KAIST

Daejeon, South Korea

Ph.D. Candidate in Artificial Intelligence

Mar. 2023 – Feb. 2027 expected

- Advisor: [Eunho Yang](#)
- Dissertation focus: Bridging Visual Manipulation to World Models.

KAIST

Daejeon, South Korea

M.S. in Artificial Intelligence

Mar. 2021 – Feb. 2023

- Advisor: [Eunho Yang](#)
- Thesis: Robust Text-guided Image Editing with StyleGAN.

Yonsei University

Seoul, South Korea

B.S. in Applied Statistics

Mar. 2017 – Feb. 2021

- Full Scholarship
- GPA: 4.03/4.5.

Invited Talks

AI EXPO KOREA KAIST AI Tech Seminar

Presenter

COEX, Seoul
2025

Service & Additional Credentials

- Reviewer service: ICML / NeurIPS / ICLR / CVPR.
- Awards: Yonsei University Full Scholarship.