

Lee, Doehyeon

Tel: +82 10 4689 0106 | E-mail: anthony16@snu.ac.kr | [Homepage](#) | [Google Scholar](#)

RESEARCH INTERESTS: *Multimodal AI Safety; Human-Aligned Safety Evaluation*

EDUCATION

MS, Industrial Engineering, Seoul National University (GPA: 4.02/4.30) Mar. 2025 – Feb. 2027 Expected

- Supervisor: Prof. [Jonghun Park](#) ([Information Management Lab](#))

BS, Industrial Engineering, Seoul National University (GPA: 3.76/4.30) Mar. 2019 – Feb. 2025

CONFERENCE PAPERS

[ICML 2025 Poster] ELITE: Enhanced Language-Image Toxicity Evaluation for Safety ([arXiv](#))

- **(Co-First Author)** Proposed **ELITE**, a benchmark of 4,587 image–text pairs for evaluating harmfulness in Vision-Language Models (VLMs), addressing insufficient harmfulness coverage and ambiguity in prior benchmarks.
- Developed a rubric-based evaluator that explicitly integrates a Toxicity Score to improve alignment with human evaluations of multimodal harmfulness.

[ICLR 2026 Poster] Jailbreaking on Text-to-Video Models via Scene Splitting Strategy ([arXiv](#))

- **(Co-author)** Contributed to **SceneSplit**, a **black-box jailbreak method** that splits unsafe narratives into sequential benign scenes to exploit safety vulnerabilities in Text-to-Video models.
- Constructed an evaluation set across 11 safety categories and defined harmfulness criteria for assessing video outputs.

CONFERENCE PRESENTATIONS

[NeurIPS 2026 Education Track] AI Safety Lives in a Direction: Reading, Breaking, and Repairing Refusal in Multimodal Models (Accepted) ([link](#))

- **(First Author)** Designed a lecture and hands-on lab teaching AI safety through the refusal direction in a vision-language model, including over-refusal checks on intervention side effects.

[NeurIPS 2026 Education Track] Can a model think VISUALLY? (Accepted) ([link](#))

- **(Co-author)** Contributed to a lecture note and Colab lab about visual reasoning in unified multimodal models.

WORK EXPERIENCE

Nanyang Technological University Global Connect Fellowship Feb. 2026 – Mar. 2026

- Collaborated at the [Data Privacy and Security Lab](#) (Supervisor: Prof. [Wei Dong](#)), **NTU Singapore**, to investigate **bias in LLMs across real-world tasks**, examining how models favor certain identities over equally valid alternatives.
- Applied sparse autoencoders to identify internal representations driving biased outputs.

AIM Intelligence ([link](#)) *Research Intern* Oct. 2024 – Feb. 2025

- Contributed to a VLM Safety Benchmark (ICML 2025) and Text-to-Video jailbreak method (ICLR 2026).
- Executed red teaming and jailbreak testing for an LLM-based care-call service, eliciting unsafe medical advice.

NAVER Cloud ([link](#)) *Intern, Hyperscale AI Planning* Jul. 2024 – Sep. 2024

- Formulated partnership strategies for online news services in the age of generative search.

Music & Audio Research Group, Seoul National University *Research Intern* Jul. 2023 – Aug. 2023

- Built a CNN-based chorus detection model (Supervisor: Prof. [Kyogu Lee](#)).

HONORS & FUNDING

- AI Seoul Tech Research Support Program, Seoul Future Foundation (Seoul Metropolitan Government), 2025 & 2026
- National Research Foundation of Korea (NRF) M.S. Research Scholarship — Lead Researcher, “AI Toolkit for Composer: Improving Text-to-Symbolic Music Generation Models via Synthetic Data,” 2025

TEACHING & SERVICE

Industrial Application of Machine Learning *Teaching Assistant* Sep. 2025 – Dec. 2025

- Graduate-level core course, Department of Industrial Engineering, Seoul National University

Basic Computing: First Adventures in Computing *Tutor* Mar. 2025 – Dec. 2025 & Sep. 2026 – Dec. 2026

- Introductory undergraduate course, Seoul National University

Reviewer, Trustworthy AI for Good (AI4GOOD) Workshop @ NeurIPS 2026 & ICML 2026