Lee, Doehyeon

Tel: +82 10 4689 0106 | E-mail: anthony16@snu.ac.kr | Address: Seoul National University | Homepage

EDUCATION

MS, Industrial Engineering, Seoul National University — 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

PUBLICATIONS

[ICML 2025 Poster] ELITE: Enhanced Language-Image Toxicity Evaluation for Safety (Arxiv)
- **(Co-First Author)** Proposed the ELITE Benchmark to overcome low harmfulness and ambiguity in existing safety evaluation benchmarks for Vision Language Models (VLMs).
- Developed the novel ELITE Evaluator, explicitly integrating a Toxicity Score to achieve superior alignment with human evaluations in accurately assessing multimodal harmfulness.
- Contributed to the development of safer, more robust VLMs by introducing the ELITE framework, setting a new standard for evaluating and mitigating safety risks in real-world applications.

[Under Review] Jailbreaking on Text-to-Video Models via Scene Splitting Strategy (Arxiv)
- **(Co-author)** Contributed to the development of SceneSplit, a novel black-box jailbreak method designed to address the significant safety gap in state-of-the-art Text-to-Video (T2V) models.
- Developed a mechanism that manipulates the generative output space by fragmenting a harmful narrative into sequential benign scenes to bypass safety filters.
- Validated on major T2V models (Luma Ray2, Hailuo, Veo2, etc.), achieving a high average Attack Success Rate (ASR) of up to 84.1%, significantly outperforming the existing baseline.

WORK EXPERIENCE

AIM Intelligence (link) *Research Intern* — Oct. 2024 – Feb. 2025
- Designed and built a VLM Safety Benchmark with improved data categorization and quality. (ICML 2025 Poster)
- Executed red teaming and jailbreak testing for a care-call service, identifying key vulnerabilities and compiling results
- Continuously monitored and analyzed AI security trends, focusing on the latest models and emerging risks

NAVER Cloud (link) *Intern, Hyperscale AI planning, Eco biz Dev* — Jul. 2024 – Sep. 2024
- Analyzed market trends and opportunities in GenAI services and AI Agents, leveraging collected data to propose strategic directions for competitive advantage
- Researched partnership strategies in key service areas, producing documentations for business development planning

Music & Audio Research Group, Seoul National University *Research Intern* — Jul. – Aug. 2023
- Supervisor: Prof Kyogu Lee (Music & Audio Research Group)
- Developed and optimized a CNN-based chorus detection model for enhanced musical structure analysis

RESEARCH GRANTS & SCHOLARSHIPS

- National Research Foundation of Korea (NRF) M.S. Research Scholarship (NRF, 2025)
  Research Topic: *AI Toolkit for Composer: Improving Text-to-Symbolic Music Generation Models via Synthetic Data*
- BK21 FOUR Graduate Research Fellow (Seoul National University, 2025)
- AI Seoul-Tech Graduate School Scholarship (Seoul Scholarship Foundation, 2025)

ADDITIONAL INFORMATION

- Language: Korean (Native), English (Fluent - TEPS score: 460 (2024)), Mandarin (beginner)
- Programming skills: Python (5-year experience), SQL (1-year experience)
- Military Service: Discharged from the Republic of Korea Army (Mar. 2021 – Sep. 2022)

RESEARCH INTEREST

- Trustworthy AI and Generative Model Safety: Focusing on vulnerability assessment, adversarial attacks (jailbreak), and ethical alignment for large-scale generative models (VLM, T2V), including the development of robust evaluation benchmarks and mitigation strategies.