Sangwoong Yoon, Ph.D.

Research Fellow in Reliable AI Alignment
at University College London (UCL)
Gower Street, London, WC1E 6BT

+82-10-2975-6589
sangwoong.yoon@ucl.ac.uk
sangwoong24yoon@gmail.com
https://github.com/swyoon

Research Interests

- Discovering and understanding new underlying principles behind generative modeling and reinforcement learning.
- Building AI agents that can interact with humans and the world safely and reliably.
- Applying AI to real-world problems, including robotics and natural sciences.

Education

Seoul National University Mar 2020 - Aug 2023
Ph.D. in Mechanical Engineering
Advisor: Frank Chongwoo Park
Thesis: Energy-Based Probabilistic Models for Epistemic Uncertainty Quantification
Outstanding Doctoral Dissertation Award

Seoul National University Mar 2014 - Feb 2016
M.S. in Interdisciplinary Program in Neuroscience
Advisor: Byoung-Tak Zhang (Department of Computer Science and Engineering)
Thesis: Adaptive Bayesian Optimization for Organic Material Screening

Hong Kong University of Science and Technology Aug 2010 - Dec 2010
Exchange Program

Seoul National University Mar 2008 - Feb 2013
B.S. in Chemical and Biological Engineering
GPA: 3.85 / 4.3 (*cum laude*)

Gyeonggi Science High School Mar 2006 - Feb 2008
Valedictorian, Top of Graduating Class

Employment

University College London, London, UK Jan 2025 - Present
Research Fellow in Reliable AI Alignment,
Department of Electronic and Electrical Engineering
Advisor: Ilija Bogunovic

- Research on reliable alignment of large language models via reinforcement learning.

Korea Institute for Advanced Study (KIAS), Seoul, Korea Sep 2023 - Jan 2025
AI Research Fellow, Center for AI and Natural Sciences

- Research on the fundamental connection between generative modeling and reinforcement learning.

Amazon.com, Seattle, WA, USA Jun 2022 - Sep 2022
Applied Scientist Intern, Search Science and AI

- Research on incorporating uncertainty information into a large language model to improve click-through rate prediction in advertisement.
- Received “inclined to hire” evaluation.

Kakao Brain, Seoul, Korea Oct 2019 - May 2020
Research Scientist Intern, Video Intelligence Team

- Research on scene-graph based image-to-image and text-to-image retrieval.

Saige Inc., Seoul, Korea Mar 2019 - Sep 2019
Researcher

- Develop deep learning-based optical defect inspection solutions for manufacturers.
- Research on deep learning algorithms for supervised and unsupervised anomaly detection.

Haezoom Inc., Seoul, Korea Jan 2016 - July 2018
Lead of machine learning team

- Led a team of five to develop a solar power forecasting system.
- Develop a data processing pipeline that integrates data from weather stations, satellites, numerical weather forecasters, and solar power plants.
- Develop fault detection system for solar power plants.
- Develop future cloud movement prediction algorithms based on 3D convolutional neural networks.

Research Visits

Heidelberg University, Heidelberg, Germany Mar 2025
Institute for Theoretical Physics
Host: Prof. Tilman Plehn

University College London, London, UK Jul 2024 - Sep 2024
Department of Electronic and Electrical Engineering
Host: Prof. Ilija Bogunovic

- Research collaboration on reinforcement learning from human feedback for large language models.

Heidelberg University, Heidelberg, Germany Mar 2023
Institute for Theoretical Physics
Host: Prof. Tilman Plehn

- Application of deep learning-based anomaly detection algorithms to high-energy physics data.

Ohio State University, Columbus, OH, USA Dec 2022
Department of Psychology
Host: Prof. Jay Myung

- Discussion on improving Bayesian optimization using Generative Gaussian Processes.

Awards

- **Outstanding Doctoral Dissertation Award** Aug 2023
  Department of Mechanical Engineering, Seoul National University
- **Qualcomm Innovation Fellowship Korea 2021**, Qualcomm Korea Sep 2021
  Awarded for "Autoencoding Under Normalization Constraints"
- **Youlchon AI Stars Scholarship 2021**, SNU AI Institute Aug 2021
- **Best Poster Award & Most Popular Poster Award** Aug 2021
  Machine Learning Summer School (MLSS) 2021 Taipei

- **Best Poster Award**, The AI KOREA 2019 Aug 2019
The first place among poster presentations
- **Cum laude**, Seoul National University Feb 2013
- **Four-year full tuition scholarship**, Korea Student Aid Foundation 2008 - 2012
- **Gyeonggi Province Governer Award**, Geyonggi Science High School Feb 2008
Awarded as the valedictorian

Publications

Preprints

1. **Sangwoong Yoon***, Himchan Hwang*, Hyeokju Jeong*, Dong Kyu Shin*, Che-Sang Park, Sehee Kweon, Frank C. Park. **Value Gradient Sampler: Sampling as Sequential Decision Making**. 2025. link
2. Seongho Son*, William Bankes*, **Sangwoong Yoon***, Shyam Sundhar Ramesh*, Xiaohang Tang, Ilija Bogunovic. **Robust Multi-Objective Decoding of Large Language Models**. 2025. link
3. Xiaohang Tang*, **Sangwoong Yoon***, Seongho Son, Huizhuo Yuan, Quanquan Gu, Ilija Bogunovic. **Game-Theoretic Regularized Self-Play Alignment of Large Language Models**. 2025. link
4. Lorenz Wolf, **Sangwoong Yoon**, Ilija Bogunovic. **This Is Your Doge, If It Please You: Exploring Deception and Robustness in Mixture of LLMs**. 2025. link

Books

1. Frank C. Park, Yonghyeon Lee, Cheongjae Jang, Seongyeon Lee, and **Sangwoong Yoon**. **Manifold, Geometry, and Machine Learning** (in preparation, expected 2025).
2. Authors: Kevin M. Lynch, Frank C. Park, Translators: Byongho Lee, **Sangwoong Yoon**, Jaewoon Kwon, Younghun Kim, Jongmin Kim, Jungbin Lim, Minjun Sohn, Jin Jung, Sanghyeon Lee, and Woosung Yang. **Modern Robotics**. Acorn Publishing, 2023 (Translation from English to Korean).
3. Daeil Kwon, Mintaek Kwon, Jungwan Mok, Geunjueong Yu, and **Sangwoong Yoon**. 과학고 공부벌레들 (**Bookworms of Science High School**). Dasan Books. 2008.

Journals

1. Woobin Yi, Dae Yeon Kim, Howon Jin[†], Sangwoong Yoon[†], and Kyung Hyun Ahn. **Early Detection of Pore Clogging in Microfluidic Systems with 3D Convolutional Neural Network**. *Separation and Purification Technology.* 2025 (in press). link IF 8.2, JCR Top 8.5%

[†] Co-correspondence

2. Shalil Khanal, Yuanhang Liu, Adebowale O. Bamidele, Alexander Q. Wixom, Alexander M. Washington, Nidhi Jalan-Sakrikar, Shawna A. Cooper, Ivan Vuckovic, Song Zhang, Jun Zhong, Kenneth L. Johnson, M. Cristine Charlesworth, Iljung Kim, Yubin Yeon, **Sangwoong Yoon**, Yung-Kyun Noh, Chady Meroueh, Abdul Aziz Timbilla, Usman Yaqoob, Jinhang Gao, Yohan Kim, Fabrice Lucien, Robert C. Huebert, Nissim Hay, Michael Simons, Vijay H. Shah, and Enis Kostallari. **Glycolysis in hepatic stellate cells coordinates fibrogenic extracellular vesicle release spatially to amplify liver fibrosis**. *Science Advances*, 2024. link IF 13.6, JCR Top 2.342%

3. Howon Jin[*], **Sangwoong Yoon**[*], Frank C. Park, and Kyung Hyun Ahn. **Data-driven constitutive model of complex fluids using recurrent neural networks**. *Rheologica Acta*, 2023. link IF 2.3, JCR Top 42.0%

4. Minwoo Lee[*], **Sangwoong Yoon**[*], Juhan Kim, Yuangang Wang, Keeman Lee, Frank Chongwoo Park, Chae Hoon Sohn. **Classification of Impinging Jet Flames Using Convolutional Neural Network with Transfer Learning**. *Journal of Mechanical Science and Technology*, 2022. link IF 1.5, JCR Top 62.8%

5. Kyu Min Park, Younghyo Park, **Sangwoong Yoon**, and Frank C. Park. **Collision Detection for Robot Manipulators Using Unsupervised Anomaly Detection Algorithms**. *IEEE Transactions on Mechatronics*, 2021. link IF 6.1, JCR Top 15.5%

Peer-Reviewed Conference Papers

1. **Sangwoong Yoon**, Himchan Hwang, Dohyun Kwon, Yung-Kyun Noh, and Frank C. Park. **Maximum Entropy Inverse Reinforcement Learning of Diffusion Models with Energy-Based Models**, *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. **Oral Presentation** (Acceptance rate: 0.46%) link

2. **Sangwoong Yoon**, Young-Uk Jin, Yung-Kyun Noh, and Frank C. Park. **Energy-Based Models for Anomaly Detection: A Manifold Diffusion Recovery Approach**, *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. link

3. **Sangwoong Yoon**, Frank C. Park, Gunsu S. Yun, Iljung Kim, and Yung-Kyun Noh. **Variational Weighting for Kernel Density Ratios**, *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. link

4. Yonghyeon Lee, **Sangwoong Yoon**, Minjun Son, and Frank C. Park. **Regularized Autoencoders for Isometric Representation Learning**, *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. link

5. **Sangwoong Yoon**, Yung-Kyun Noh, and Frank C. Park. **Autoencoding Under Normalization Constraints**, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. link

[*] Equal contribution

6. **Sangwoong Yoon**, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. **Image-to-Image Retrieval by Learning Similarity between Scene Graphs**, *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021. link

7. SooKyung Kim, Hyojin Kim, Joonseok Lee, **Sangwoong Yoon**, Samira E. Kahou, Karthik Kashinath, Mr Prabhat. **Deep Hurricane-Tracker: Tracking and Forecasting Extreme Climate Events**, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. link

Workshop Papers

1. **Sangwoong Yoon***, William Bankes*, Seongho Son*, Anja Petrovic*, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. **Group Robust Best-of-K Decoding of Language Models for Pluralistic Alignment**. *Neural Information Processing Systems 2024 Pluralistic Alignment Workshop*, 2024.

2. **Sangwoong Yoon**, Frank C. Park, and Yung-Kyun Noh. **Kullback-Leibler Divergence Estimation using Variationally Weighted Kernel Density Estimators**, *Neural Information Processing Systems 2019 Information Theory and Machine Learning Workshop*, 2019.

3. **Sangwoong Yoon**, Yonho Song, Minsoo Kim, Frank C. Park and Yung-Kyun Noh. **Interpretable Feature Selection Using Local Information for Credit Assessment**. *Neural Information Processing Systems 2018 Workshop on Challenges and Opportunities for AI in Financial Services*, 2018. **Oral Presentation**

4. **Sangwoong Yoon**, Sang-Woo Lee, and Byoung-Tak Zhang, **Predictive Property of Hidden Representations in Recurrent Neural Network Language Models**, *Neural Information Processing Workshop Systems 2014 Workshop on Modern Machine Learning Methods and Natural Language Processing*, 2014.

Patents

1. Oh-Hyun Kwon, Jung-Seok Hyung and **Sangwoong Yoon**, **Method, Server, and System for Detecting Abnormality of a Power Plant using Solar Energy**, the Republic of Korea patent, KR101775065B1, applied in Aug 5, 2016, granted in Aug 30, 2017.

2. Oh-Hyun Kwon, Jung-Seok Hyung and **Sangwoong Yoon**, **Method and Server for Forecasting Generation of a Power Plant using Solar Energy**, the Republic of Korea patent, KR101808047B1, applied in Aug 5, 2016, granted in Dec 14, 2017.

Invited Talks

International

- **University of Cambridge**, Cambrdige, UK — May 2025
- **University of Oxford**, Oxford, UK — Apr 2025
  Peierls Rodulf Centre for Theoretical Physics
  Title: Generative Modeling is Imitation Learning

- **Heidelberg University**, Heidelberg, Germany Mar 2025
  Institute for Theoretical Physics (Host: Tilman Plehn)
  Title: My Research Journey: From Anomaly Detection To Inverse Reinforcement Learning
- **Imperial College London**, London, UK Feb 2025
  CSML Reading Group (Host: Yingzhen Li)
  Title: Generative Modeling is Imitation Learning and Sampling is Reinforcement Learning
- **University of Oxford**, Oxford, UK Feb 2025
  Oxford Robotics Institute (ORI)
  Title: Generative Modeling is Imitation Learning
- **RIKEN-AIP**, Tokyo, Japan Dec 2024
  The 87th TrustML Young Scientist Seminar (Host: Masashi Sugiyama)
  Title: Sampling is Reinforcement Learning and Generative Modeling is Imitation Learning
- **University of Michigan**, Ann Arbor, USA Oct 2024
  Electrical Engineering and Computer Science (Host: Stella X. Yu)
  Title: Maximum Entropy Inverse Reinforcement Learning of Diffusion Models with Energy-Based Models
- **University College London**, London, UK Jul 2024
  Department of Electronic and Electrical Engineering (Host: Ilija Bogunovic)
  Title: Maximum Entropy Inverse Reinforcement Learning of Diffusion Models with Energy-Based Models
- **University of Cambridge**, Cambridge, UK Feb 2024
  Department of Applied Mathematics and Theoretical Physics (Host: Carola-Bibiane Schönlieb)
  Title: Why autoencoders fail at anomaly detection and what we can do about it
- **Heidelberg University**, Heidelberg, Germany Mar 2023
  Institute for Theoretical Physics (Host: Tilman Plehn)
  Title: Rethinking autoencoder-based anomaly detection from probabilistic perspective
- **Ohio State University**, Columbus, USA Dec 2022
  Department of Psychology (Host: Jay Myung)
  Title: Gaussian processes are density estimators

Domestic

- **Ulsan National Institute of Science & Technology** Apr 2025
  AIGS Seminar
  Title: Reinforcement Learning for Non-Reinforcement Learning Problems
- **Focused Workshop on AI in High Energy Physics** Jan 2025
  Title: Generative Modeling is Imitation Learning
- **Department of Biological Sciences, Seoul National University** Dec 2024
  Title: Sampling is Reinforcement Learning and Generative Modeling is Imitation Learning

- **APCTP-SISSA Joint Workshop on AI and Theoretical Physics** Dec 2024
  Title: Sampling is Reinforcement Learning and Generative Modeling is Imitation Learning
- **Saige Inc.** Oct 2024
  Title: Energy-Based Models for Classifying In-and-Out
- **The Korean Institute of Chemical Engineers Fall Meeting** Oct 2024
  Title: AI in Manufacturing: Will Revolution Come?
- **Innovation Center for Industrial Mathematics, National Institute for Mathematical Sciences** May 2024
  Title: Diffusion by Dynamic Programming
- **Korean Mathematical Society Spring Meeting 2024** Apr 2024
  Title: Maximum Entropy Inverse Reinforcement Learning of Diffusion Models with Energy-Based Models
- **Korea Research Institute of Chemical Technology** Feb 2024
  Title: Training Diffusion Models with (Inverse) Reinforcement Learning
- **KCMS-Theory Workshop** Dec 2023
  Title: Why autoencoders fail at anomaly detection and what we can do about it
- **College of Agriculture and Life Sciences** Dec 2023
  Seoul National University
  Title: Why autoencoders fail at anomaly detection and what we can do about it
- **High-Energy Physics and AI Workshop, Hanyang University** Dec 2023
  Title: Why autoencoders fail at anomaly detection and what we can do about it
- **Robot Intelligence Lab, Korea University** Nov 2023
  Title: Generative Modeling is Imitation Learning
- **AI and Quantum Information for Particle Physics, KAIST** Nov 2023
  Title: Why autoencoders fail at anomaly detection and what we can do about it
- **IITP Workshop on Video Understanding and Generation using Knowledge-based Deep Logic Neural Networks** Sep 2023
  Title: Energy-Based Models for Classifying In and Out
- **Data Science Career Day, Graduate School of Data Science**, Sep 2023
  Seoul National University
  Title: Lessons from Three Degrees from Three Departments
- **LG AI Research** Feb 2022
  Title: Autoencoding Under Normalization Constraints

Grants

- **Developing Reliable Foundation Models with Theoretical Framework and Scalable Personalization** Aug 2024 - Aug 2027
  Ministry of Science and ICT Global Basic Research Laboratory
  PI: Hye Won Chung (KAIST)
  Role: Participating researcher

- **Investigation on Theoretical Connection between Generative Modeling and Reinforcement Learning** Sep 2023 - Aug 2025

KIAS Basic Research Grant
PI: Sangwoong Yoon

- **Development of Training and Inference Methods for Goal-Oriented Artificial Intelligence Agents** Apr 2022 - Dec 2026
  IITP Human-Centric AI Core Technology Development
  PI: Frank Chongwoo Park (SNU)
  Role: Lead author of the proposal and main researcher

- **LIDAR-Based Lane Detection**, Seoul Robotics Jun 2022 - Dec 2022
  PI: Frank Chongwoo Park (SNU)
  Role: Lead author of the proposal and main researcher

- **Development of a Machine Learning-Based Solution for Anomaly Detection and Root Cause Diagnosis in Solar Power Generation Using Meteorological and Power Monitoring Data** Jun 2016 - Jul 2017
  Small and Medium Business Administration
  PI: Oh-Hyun Kwon (Haezoom Inc.)
  Role: Lead author of the proposal and main researcher

- **Development of Method for Accelerating Organic Material Search using Machine Learning** Apr 2014 - Apr 2015
  Samsung Advanced Institute of Technology
  PI: Byong-Tak Zhang (SNU)
  Role: Lead author of the proposal and main researcher

Teaching

- **Time-Series Forecasting Tutorial**, SK Telecom Apr 2024
  Instructed a 3-hour tutorial on time-series forecasting using deep learning methods.

- **KIAS-Hanyang AI Summer School**, Hanynag University Oct 2023
  Instructed two 3-hour lectures: “Introduction to DDPM” and “Diffusion Model Hands-on Tutorial.”

- **Guest Lecture on Information Geometry**, Seoul National University Nov 2022
  Delivered a guest lecture in the course Geometric Methods for High-Dimensional Data Analysis, taught by Prof. Frank Park.

- **Introduction to Machine Learning**, Microrheology Laboratory Aug 2020
  Department of Chemical and Biological Engineering, Seoul National University
  Instructed a 20-hour course on machine learning and deep learning, including coding practice sessions.

- **Interpretable Machine Learning Course**, Fast Campus Apr 2019
  One-day lecture on interpretable machine learning

- **Variational Autoencoder Course**, Fast Campus Apr 2018
  Two-day lecture on variational autoencoders

Professional Services

Services for Academic Communities

- Area chair of NeurIPS 2025 Mar 2025
- Reviewer of NeurIPS, ICML, ICLR, AAAI, CVPR, ICCV, AISTATS, and ACML 2019 - Present
- Co-organizer of KIAS-Hanyang AI Summer School Oct 2023
- Organizer of IITP Joint Workshop between Frank Park's project and Byong-Tak Zhang's project Sep 2023
- Website admin for Korea-Japan Machine Learning Workshop 2019 Feb 2019

Services for Developer Communities

- Contributor of Pandas, an open-source Python library: Submitted 5 merged pull requests to Pandas: #17253, #19427, #22380, #26157, #26158
- Staff of PYCON KR 2015 and PYCON APAC 2017

Media Coverage

- 고등과학원, 새로운 생성 AI 분야 알고리즘 제시, 전자신문, 2024-12-05. link
- 생성형 AI에 모방학습 적용 알고리즘 개발...속도 10배 높여, 연합뉴스, 2024-12-05. link

References

- **Ilija Bogunovic** (`i.bogunovic@ucl.ac.uk`)
  Lecturer, Department of Electronic and Electrical Engineering,
  University College London (Postdoc Advisor)
- **Frank Chongwoo Park** (`fcp@snu.ac.kr`)
  Professor, Department of Mechanical Engineering,
  Seoul National University (Ph.D. Advisor)
- **Yung-Kyun Noh** (`nohyung@hanyang.ac.kr`)
  Associate Professor, Department of Computer Science, Hanyang University
- **Hyokun Yun** (`yunhyoku@amazon.com`)
  Principal Applied Scientist, Amazon.com (Internship Manager)
- **Tilman Plehn** (`plehn@thphys.uni-heidelberg.de`)
  Professor, Institute for Theoretical Physics, Heidelberg University