Sanghyun Yi

San Jose CA 95134, USA
syi@caltech.edu • +1 408 707 8856 • https://www.sanghyunyi.com

ABOUT

PhD in Computational Neuroscience specializing in Machine Learning and Generative AI. Expertise in optimal arbitration in mixture of experts and reinforcement learning, with multiple publications and patents. Industry experience includes optimizing video diffusion models via quantization at Samsung and developing retrieval-augmented generation (RAG) systems for multi-hop QA at Amazon. Skilled in post-training of generative models, deep learning architectures, and efficient inference techniques.

EDUCATION

PhD, **Caltech**, Pasadena, USA — Sep 2018 – Sep 2024

- Social and Decision Neuroscience
- Emphasis on Computational Cognitive Neuroscience and Machine Learning
- Supervisor: John O'Doherty
- Chen graduate fellow

BS, **Seoul National Univeristy**, Seoul, Korea — Mar 2011 – Feb 2018

- College of Liberal Studies
- Major: Mathematics
- Minor: Computer Science, Statistics
- Cum laude
- Mandatory military service from 2013 to 2015

SKILLS

LLM, RAG, Agentic AI, NLP, Machine Learning, Deep Learning, Statistical Analysis, Programming (Python, PyTorch, TensorFlow, Scikit-learn, NumPy, Scipy, Pandas, MATLAB, R, SQL, Unix/Linux), Diffusion Model, Quantization, Experimental Design, Research, Problem Solving, Communication, Collaboration and more.

EXPERIENCE

Amazon Web Services (AWS), Amazon — Sep 2024 – Jan 2025

- Applied Scientist Intern (Applied Scientist II)
- Proposed a agentic **RAG** system for solving **multi-hop QA** problems, utilizing modularized reasoning, sequence-level-distillation and post-training optimization (e.g., Kahneman-Tversky Optimization) for improved performance.
- Demonstrated significant increase in accuracy, exact match, F1, and retrieval recall across 5 public benchmark datasets compared to foundation model, ReACT, and other existing methods.
- Currently working on a paper.

AI systems, Samsung Semiconductor, Inc. — Jun 2024 – Sep 2024

- AI Systems Research Scientist Intern
- Developed hardware-friendly static **quantization methods for video diffusion models**. The methods achieved the performance of the current best method for quantizing video diffusion transformers which uses dynamic quantization. The paper has been submitted, and the patent is pending.

Alexa AI, Amazon Lab126 — Mar 2018 – Aug 2018

- Applied Scientist Intern
- Developed **automatic conversation evaluators** using the transformer sentence embedding and a range of natural language features and applied the evaluators in **improving end-to-end language generation models** which is **the early form of LLM with RLHF**. The conversation evaluator models were deployed to worldwide Alexa Prize competitors to automate the assessment of their chatbot outputs. The proposed models and their application were presented at the peer-reviewed conference (NAACL NeuralGen 2019), published in the peer-reviewed proceedings (INLG 2019) and cited in numerous papers including the RLHF papers by **OpenAI**. This framework has been officially patented in the US (US11194973B1). I received a return offer but declined it in order to focus on my graduate research.

O'Doherty Lab, Caltech Sep 2018 – Sep 2024

- Graduate Research Assistant
- Developed a **CNN model that predicts human aesthetic ratings on art paintings** with a correlation accuracy of 0.4. The CNN was used to model human aesthetic preference and its neural mechanisms. By using decoding analyses, unveiled both the CNN and the human brain construct the aesthetic preferences through hierarchical computation of various visual features.
- Developed a **human-visual-stream-inspired CNN model** that predicts action affordances, or the most suitable actions, on everyday objects. Its accuracy reached 66% while the level of noise ceiling of the ground truth data was 68%. Proposed a weakly-supervised affordance segmentation method using the saliency maps produced by the guided backpropagation on the trained CNN model. The human evaluation on the segmentation results showed that the proposed method achieves 42% accuracy, which is significantly higher than the chance level of 25%.
- Developed **novel MOE reinforcement learning frameworks** that best explains the neuro-computational mechanism of the learning and decision-making in humans than the previously proposed models.
- Developed a **real-time hand gesture tracking system using Openpose, AWS and a deep learning classifier** with a classification accuracy 93% to implement an ecologically valid behavioral experiment.
- Developed **machine-learning-based tool for transcribing movie and extracting natural language features**, including semantics, sentiments, part-of-speech, and dialogue act. This tool was instrumental in analyzing brain activities during movie watching.
- **Designed and implemented behavioral/fMRI experiments** to test human learning in various conditions (e.g., contrasting between when previous knowledge or bias either hinders or fosters reward-related learning).
- Conducted **statistical analyses of human choice behavior and fMRI data** which involved Bayesian model comparison, statistical inference, GLM analyses and machine learning based decoding analyses.

Laboratory for Brain and Machine Intelligence, KAIST Jun 2017 – Feb 2018

- Undergraduate Research Intern
- Developed a **DDQN algorithm that dynamically control the behavioral experiment** that guide human behavior in a desired way. The research was patented in Korea (10-2018-0089185), with a pending patent application in the US (16381954)
- Developed an **AI soccer robot algorithm** and achieved 3rd place in AI World Cup 2017

Machine Intelligence Lab, Seoul National University Jan 2016 – Dec 2016

- Undergraduate Research Intern
- Designed a **bilingual sentence alignment algorithm** based on length, word representational vectors and dictionary information.
- Proposed a **CNN based encoder** in attention based encoder decoder model.

Natural Language Processing Group, MIT Jun 2011 – Aug 2011

- Undergraduate Research Intern
- **Visualized the co-authorship network** of MIT faculty members using the data mining and web scrapping

ORGANIZATIONS &ACTIVITIES

Caltech Brain Imaging Center Meeting, Caltech Apr 2022 – Sep 2022

- Organizer
- I coordinated and managed a biweekly talk series at the Caltech Brain Imaging Center (CBIC), fostering engagement within the Caltech brain imaging community. The series showcased current CBIC research projects and delved into advanced topics in brain imaging methodologies.

Caltech Korean Graduate Student Association, Caltech Feb 2021 – Feb 2022

- President
- I spearheaded the planning and coordination of social events, fostering a sense of community among Korean graduate students and postdocs at Caltech while also extending our reach to include the Korean communities at UCLA and USC.

259 Company, Gangbuk Police, Seoul Metropolitan Police Agency Oct 2013 – Jul 2015

- Company Commander (Sergeant)
- Mandatory military service.
- I led a company of about 100 constables.
- I was awarded commendations from the Commissioner of Seoul Metropolitan Police Agency and the Senior Superintendent of Gangbuk Police for my achievements during the service.

HONORS &AWARDS

- **Chen Graduate Innovator Grant Awards**, Chen Institute for Neuroscience, Caltech, Jan 2022
  Awarded $10,000 for my proposal, Behavioral and Neural Understanding of Affordance in Value-based Decision-Making.
- **National Science Foundation Graduate Research Fellowship**, National Science Foundation, 2020
  Honorable Mention
- **Chen Graduate Fellow**, Chen Institute for Neuroscience, Caltech 2018 – 2019
- **National Scholarship for Science and Engineering**, Korea Student Aid Foundation 2011 – 2016
  Full tuition & fee scholarship for my undergraduate studies due to my outstanding academic performance.
- **Semifinalist, The Alexa Prize** Aug 2017
  Developed a social bot for The Alexa Prize which was serviced to the entire Alexa user base.
  10th place among 18 semifinalists which include 12 sponsored teams.
  I was the de facto leader of the team, which was unsponsored and was the only semifinalist team from Asia.
- **3rd place, Nvidia Deep Learning Contest** Oct 2016
  Achieved 85.1% accuracy on food image classification using the Inception v3.
  The only undergraduate awardee.
- **3rd place, AI World Cup 2017** Dec 2017
  Developed an AI soccer robot.
  1st place at the preliminary league and 3rd place at the final tournament where the top 4 teams of the preliminary participated in. Awarded approx $1500.
  Gave a talk about the result at The 5th International Conference on Robot Intelligence Technology and Applications(RiTA).

SELECTED PUBLICATIONS

[1] Sanghyun Yi, Qingfeng Liu, and Mostafa El-Khamy, “Hardware-Friendly Static Quantization Method for Video Diffusion Transformers”, in *arXiv*:2502.15077, 2025.

[2] Sanghyun Yi and John P. O’Doherty, “Computational and neural mechanisms underlying the influence of action affordances on value learning”, in *BioRxiv* (revision submitted) 2023.

[3] Kiyohito Iigaya, Sanghyun Yi, Iman Wahle, Sandy Tanwisuth, Logan Cross and John P. O’Doherty, “Neural mechanisms underlying the hierarchical construction of perceived aesthetic value”, in *Nature Communications* 14 (1), 127 2023.

[4] Kiyohito Iigaya, Sanghyun Yi, Iman Wahle, Sandy Tanwisuth and John P. O’Doherty, “Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features”, in *Nature Human Behaviour* 5, 743-755 2021.

[5] Kiyohito Iigaya, Sanghyun Yi, Iman Wahle, Sandy Tanwisuth, Aniek Fransen and John P. O’Doherty, “Computational principles of value construction”, in *Computational and Systems Neuroscience (COSYNE)* 2021.

[6] Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel and Dilek Hakkani-Tur, “Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators”, in *International Conference on Natural Language Generation (INLG)*, Tokyo, Japan, 2019. (Oral presentation)

[7] Sanghyun Yi, Jeehang Lee, Changhwa Lee, Juno Kim, Sujin An and Sang Wan Lee, “A Competitive Path to Build Artificial Football Agents for AI Worldcup’’, in *IEEE/IEIE International Conference on Consumer Electronics (ICCE) Asia*, Jeju, Korea, 2018.

[8] Sanghyun Yi, Jeehang Lee and Sang Wan Lee, “Maximally separating and correlating model-based and model-free reinforcement learning”, in *Computational and Systems Neuroscience (COSYNE)*, Denver, USA, 2018.

[9] Sanghyun Yi and Kyomin Jung, “A Chatbot by Combining Finite State Machine, Information Retrieval, and Bot-Initiative Strategy”, in *1st Proceedings of Alexa Prize (Alexa Prize 2017)*, Las Vegas, USA, 2017.

PATENTS

[1] Rahul Goel, Chandra Prakash Khatri, Tagyoung Chung, Raefer Christopher Gabriel, Anushree Venkatesh, Behnam Hedayatina, Sanghyun Yi, "Dialog Response Generation", US patent (11194973)

[2] Sang Wan Lee, JeeHang Lee, Sanghyun Yi, "Apparatus and method for eliciting optimal strategy of the humans in the interactive games using artificial intelligence", US patent pending (16381954)

[3] Sang Wan Lee, JeeHang Lee, Sanghyun Yi, "Apparatus and method for eliciting optimal strategy of the humans in the interactive games using artificial intelligence", Korean patent (10-2018-0089185)

TEACHING

- TA for Psy 13:Introduction to Cognitive Neuroscience — Spring 2020, 2021, 2022, 2023
- TA for EC/ACM/CS 112:Bayesian Statistics — Winter 2020, 2021, 2022

LANGUAGES

- English: fluent.
- Korean: native language.

[*CV compiled on 2025-03-29*]