Hyesu Lim

hyesulim.ac@gmail.com | hyesulim.github.io | linkedin.com/in/hyesulim | google scholar

Research Interest

My research focuses on building AI systems that are **testable, explainable, and controllable** — so that people can understand, trust, and interact with them in real-world use. I develop **interpretability** tools, especially sparse autoencoders, to expose the concepts that deep vision, language, and multimodal models learn [C07, C08, C09, W02], and build on these insights to make model representations more controllable and robust under distribution shift [C01, C02, C06]. I apply these methods across scientific domains, from medical imaging to language.

Education

Korea Advanced Institute of Science and Technology (KAIST) *Seongnam, South Korea*

MS. & Ph.D., Artificial Intelligence *Mar 2021 - Feb 2026*

- Advisor: Jaegul Choo
- GPA: 3.95/4.30
- Thesis: Trustworthy AI in the Wild: Robustness and Interpretability for Enhanced Reliability (PDF)

Korea University *Seoul, South Korea*

BS., Computer Science and Engineering *Mar 2017 - Feb 2021*

- ***Highest Honors.*** Valedictorian of College of Informatics
- GPA: 4.35/4.50

Visiting & Exchange Programs

Carnegie Mellon University *Pittsburgh, U.S.A.*

Visiting scholar, Software and Societal Systems, School of Computer Science *Aug 2023 - Feb 2024*

- Related publication: Towards Calibrated Robust Fine-Tuning of Vision-Language Models (**NeurIPS 2024**)

University of Toronto *Toronto, Canada*

Exchange undergraduate student, Computer Science *Aug 2019 - Apr 2020*

- GPA: 3.81/4.00

Research Experience

Helmholtz Munich *Munich, Germany*

Postdoctoral researcher *Mar 2026 -*

- Supervisor: Steffen Schneider
- Team: Dynamical Inference Lab, Institute of Computational Biology, Computational Health Center

NAVER AI Lab *Seongnam, South Korea*

Research Intern *Oct 2024 - Apr 2025*

- Team: Backbone Research
- Project: Mechanistic Interpretability for on Vision-Language Models
- Related publication: VisualScratchpad: Grounding Visual Concepts in Large Vision Language Models (**ICLR Workshop 2026**)

Helmholtz Munich *Munich, Germany*

Visiting Researcher *Jul 2024 - Sep 2024*

- Team: Dynamical Inference Lab, Institute of Computational Biology, Computational Health Center
- Project: Mechanistic Interpretability for Adaptation on Vision-Language Foundation Models
- Related publication: Sparse autoencoders reveal selective remapping of visual concepts during adaptation (**ICLR 2025**)

Qualcomm AI Research *Seoul, South Korea*

Interim Engineering Intern *Feb 2022 - Oct 2022*

- Project: Test-time adaptation
- Related publication: TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation (**ICLR 2023**)

Publications

International Conference Publications (*: equal contributions)

C09 **ConceptScope: Characterizing Dataset Bias via Disentangled Visual Concepts** [PDF] [Code] [Website]

Jinho Choi, **Hyesu Lim**, Steffen Schneider, and Jaegul Choo

The Thirty-Ninth Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2025.

TL;DR: We present ConceptScope, a scalable and automated framework for analyzing visual datasets by discovering and quantifying human-interpretable concepts using Sparse Autoencoders trained on representations from vision foundation models.

C08 **CytoSAE: Interpretable Cell Embeddings for Hematology** [PDF] [Code]

Muhammed Furkan Dasdelen, **Hyesu Lim**, Michele Buck, Katharina Götze, Carsten Marr, and Steffen Schneider

Medical Image Computing and Computer Assisted Intervention (**MICCAI**), 2025.

TL;DR: We propose CytoSAE, a sparse autoencoder which is trained for peripheral blood single-cell images, and show that it generalizes to bone marrow cytology, identifying morphologically relevant concepts.

C07 **Sparse autoencoders reveal selective remapping of visual concepts during adaptation** [PDF] [Code] [Poster, Slide & Video]

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider

International Conference on Learning Representations (**ICLR**), 2025.

TL;DR: Here we develop a new Sparse Autoencoder (SAE) for the CLIP vision transformer, named PatchSAE, to extract interpretable concepts at granular levels (e.g., shape, color, or semantics of an object) and their patch-wise spatial attributions.

C06 **Towards Calibrated Robust Fine-Tuning of Vision-Language Models** [PDF] [Code] [Poster, Slide & Video]

Changdae Oh*, **Hyesu Lim***, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song

Neural Information Processing Systems (**NeurIPS**), 2024.

TL;DR: This work proposes a robust fine-tuning method that improves both OOD accuracy and calibration error in Vision Language Models (VLMs) via fine-tuning with a constrained multimodal contrastive loss that minimizes a shared upper bound of OOD generalization and calibration error.

C05 **Translation Deserves Better: Analyzing Translation Artifacts in Cross-lingual Visual Question Answering** [PDF][Poster]

ChaeHun Park, Koanho Lee, **Hyesu Lim**, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo

Association for Computational Linguistics (**ACL**) Findings, 2024.

TL;DR: Our analysis reveals that translated texts contain unique characteristics distinct from human-written ones, referred to as translation artifacts. We find that these artifacts can significantly affect the models, confirmed by extensive experiments across diverse models, languages, and translation processes.

C04 **Slice and Conquer: A Planar-to-3D Framework for Efficient Interactive Segmentation of Volumetric Images** [PDF]

Wonwoo Cho*, Dongmin Choi*, **Hyesu Lim***, Jinho Choi, Saemee Choi, Hyunseok Min, Sungbin Lim, and Jaegul Choo

Winter Conference on Applications of Computer Vision (**WACV**), 2024.

TL;DR: We propose to conduct two consecutive stages for 3D interactive segmentation: 1) 2D interactive segmentation and 2) guided 3D segmentation, where 1) predicts 2D masks given user interactions and 2) predicts the final 3D output given the 2D masks.

C03 **PRiSM: Enhancing Low-Resource Document-Level Relation Extraction with Relation-Aware Score Calibration** [PDF]

Minseok Choi, **Hyesu Lim**, and Jaegul Choo

International Joint Conference on Natural Language Processing and the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (**IJCNLP-AACL**), 2023.

TL;DR: We propose the Pair Relation Similarity Module (PRiSM), which considers relation semantics by computing similarity scores between the entity pair embedding and the relation embedding, encouraging the two to stay close in the semantic space if they are related.

C02 **TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation** [PDF] [Poster, Slide & Video] [Website]

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi

International Conference on Learning Representations (**ICLR**), 2023.

TL;DR: We propose a test-time batch normalization method, which interpolates source and current batch statistics considering each layer's domain-shift sensitivity level that shows robust performance over various realistic evaluation scenarios.

C01 **AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain** [PDF] [Code] [Poster] [Slide & Video]

Jimin Hong*, Taehee Kim*, **Hyesu Lim***, and Jaegul Choo

Empirical Methods in Natural Language Processing (**EMNLP**), 2021.

TL;DR: We propose to consider a vocabulary of a pre-trained language model as an optimizable parameter, allowing us to update the vocabulary by expanding it with domain-specific vocabulary based on tokenization statistic.

International Conference Workshop Publications

W02 **VisualScratchpad: Grounding Visual Concepts in Large Vision Language Models** [PDF]

Hyesu Lim, Jinho Choi, Taekyung Kim, Byeongho Huh, Jaegul Choo, and Dongyoon Han

ICLR 2026 Workshop on Trustworthy AI (**Trustworthy AI**), 2026.

TL;DR: We introduce an interactive tool named VisualScratchpad that works with vision-language models, linking the resulting visual concepts to text tokens via text-to-image attention and allowing us to analyze which visual concepts are captured by the vision encoder and used by the language model.

W01 **Towards Calibrated Robust Fine-Tuning of Vision-Language Models** [PDF]

Changdae Oh, Mijoo Kim, **Hyesu Lim**, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Son

NeurIPS 2023 Workshop on Distribution Shifts (**DistShift**), 2023.

TL;DR: We initiate the investigation on the calibration of VLM after fine-tuning under distribution shifts and introduce simple yet effective approaches to improve calibration error.

Domestic Conference Publications (*: equal contributions)

- **A Survey on Interactive Image Segmentation Using Deep Learning** [PDF]

  Gyuhyeon Sim*, Jinho Choi*, **Hyesu Lim**, and Jaegul Choo

  Conference of Korean Artificial Intelligence Association (CKAIA 2020). p.44-48.

  TL;DR: We briefly introduce the recent literature on interactive image segmentation using deep learning techniques, grouping them based on different types of user interactions.

Granted Patents

P03 **Adapting machine learning models for domain-shifted data** [PDF]

Hyesu Lim, Byeonggeun Kim, and Sungha Choi

U.S. Patent No. US20240119360A1, 11 April, 2024.

P02 **Suggesting a New and Easier System Function by Detecting User's Action Sequences** [PDF]

Sungrack Yun, Hyoungwoo PARK, Seunghan YANG, **Hyesu LIM**, Taekyung Kim, Jaewon Choi, and Kyu Woong Hwang

U.S. Patent No. US20240045782A1, 08 Feb, 2024.

P01 **Method and Device for Proving Oral Health Care Service in Periodontal and Peri-implant Diseases** [PDF]

Shin-Young Park, Shin Hye Chung, Hayoung Kim, Dayoung Kim, **Hyesu Lim**

KR Patent No. 1024127030000, 21 Jun, 2022.

Honors, Awards, and Scholarships

2024.07 - 2024.08	**Summer Institute Programme 2024 Scholarship (EUR 3.5K),** NRF/DAAD
2023.08 - 2024.02	**Sponsored AI intensive program at CMU (USD 41K),** IITP and Sogang University
2022.08	**Qualcomm IT Tour**, Qualcomm Korea
2022.01	**Best Poster Awards 1st place**, KAIST AI Workshop 21/22
2021.03 - 2026.02	**KAIST scholarship (USD 37K; USD 3.7K per semester)**, KAIST
2019.07 - 2020.12	**Talent Development scholarship (USD 3.5K),** Chungcheongnam-do Human Resources Development Foundation
2018.03 - 2020.12	**Dooeul scholarship (USD 27K),** Dooeul

Invited Talks

- **Trustworthy AI in the Wild: Robustness and Interpretability for Enhanced Reliability** [Slide]

  Data Science for Humanity Group at Max Planck Institute for Security and Privacy (2025.06.25)

Academic Services

Reviewer

- 2026 ICLR, ECCV
- 2025 NeurIPS, ICML Workshop (PUT), NN, AAAI
- 2024 CVPR

Organizer

- Women in AI/EE/CS at KAIST 2024
- KAIST AI Workshop 21/22

Conference Volunteer

- 2025 MICCAI

Skills

Programming	Python & PyTorch (Advanced: 6+ years)
Languages	Korean (Native), English (Fluent: TOEFL iBT 110/120 (earned in 2018))
Soft Skills	Documentation, Presentation, Time Management

References

Jaegul Choo *Seongnam, South Korea*

Ph.D. Advisor (KAIST AI) *2021.03 - 2026.02*

- E-mail: jchoo@kaist.ac.kr
- website

Sungha Choi *Seoul, South Korea*

Team manager (Qualcomm AI Research) *2022.02 - 2022.10*

- E-mail: belle79@gmail.com
- website

Steffen Schneider *Munich, Germany*

Advisor (Helmholtz Munich) *2024.01 -*

- E-mail: stes@hey.com
- website

Dongyoon Han *Seongnam, South Korea*

Mentor (NAVER AI Lab) *2024.10 - 2025.04*

- E-mail: dongyoon.han@navercorp.com
- website