

Alec Helbling

Interpretability and Safety

I develop novel interfaces for controlling Generative AI. In particular, I leverage a combination of machine learning and interactive visualization to allow humans to control generative models.

I have worked with researchers and engineers at Georgia Tech, Apple, Adobe, IBM Research, NASA Jet Propulsion Laboratory, Microsoft, and the University of Pittsburgh.

✉ alechelbling1@gmail.com

🏠 alechelbling.com

📄 CV PDF

🐦 @alec_helbling

🌀 @helblazer811

🎓 Google Scholar

Education

Aug. 2023 - Present

Ph.D. in Machine Learning

Georgia Institute of Technology, Atlanta, GA

Advisor: Duen Horng (Polo) Chau, Co-advisor: None

🏆 NSF Fellowship, Georgia Tech President's Fellowship

Aug. 2018 - Dec.
2022

B.S. in Computer Science

Georgia Institute of Technology, Atlanta, GA

Overall GPA: 3.91/4.00, Highest Honors, Threads in Intelligence and Theory

Industry Research Experience

Summer 2025

Apple, Seattle, WA

AI Research Intern, Responsible AI Team,

Mentor: Joseph Cheng, Leon Gatys, Shruti Palaskar, Charlie Maalouf

I worked on improving the safety of vision language models.

Summer 2024

Adobe Firefly, San Jose, CA

AI Research Intern,

Mentor: Siddharth Iyer, Oliver Brdiczka

I worked on improving the safety of large scale diffusion models.

Summer 2021, 2022

IBM Research, Yorktown Heights, NY

AI Research Intern, Foundations of AI Reasoning Group

Mentor: Achille Fokoue, Tengfei Ma, Maria Chang

I worked on systems for summarizing natural language documents in a way that preserves their factuality.

Summer 2020

Microsoft, Redmond, WA

Software Engineering Intern, Microsoft Intune Insights and Analytics Group

Mentor: Durga Kumar Varanasi

Developed a system for automatically distributing service traffic across several database instances.

Summer 2019

NASA Jet Propulsion Lab, Pasadena, CA

Machine Learning Research Intern, Machine Learning and Instrument Autonomy Group

Mentor: Lukas Mandrake, Jack Lightholder

I worked on interactive interfaces allowing mission scientists without ML expertise to deploy powerful ML methods.

Academic Research Experience

Present —

Georgia Institute of Technology, Atlanta, GA

Graduate Research Assistant, School of Computational Science and Engineering

Advisor: Duen Horng (Polo) Chau

I am a member of the Polo Club of Data Science. We focus on the intersection of machine learning, human computer interaction, and visualization.

Fall 2020–Spring
2023

Georgia Institute of Technology, Atlanta, GA

Student Researcher, SIPLab

Mentor: Chris Rozell, Kion Fallah, Matthew O'Shaughnessy

I worked on systems for controlling deep generative models for image generation by using simple pairwise feedback from people.

Mar. 2015 – Aug.
2018

University of Pittsburgh, School of Medicine, Pittsburgh, PA

Student Researcher and Javascript Programmer, Koes Lab

Mentor: David Koes

I worked on methods for visualizing molecular structures in web browsers, applications of machine learning to computational drug discovery, and visualizations of machine learning models for drug discovery.

Honors and Awards

2024

National Science Foundation Graduate Research Fellowship

Prestigious fellowship with 3 years guaranteed funding.

2025

Oral Presentation, ICML 2025

For "ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features"

2025

Best Paper Award, CVPR Workshop on Visual Concepts

For "ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features"

2025

Spotlight Oral Presentation, CVPR Workshop on Visual Concepts

For "ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features"

2024

Best Poster, IEEE VIS 2024

For "Transformer Explainer – Interactive Learning of Text-Generative Models"

2023

Best Poster, IEEE VIS 2023

For "Manim ML: Communicating Machine Learning Architectures with Animation"

2023 — Present

President's Fellowship at Georgia Institute of Technology

Select number of 1st year doctoral students who bring exemplary levels of scholarship and innovation to their academic departments

2020, 2021

President's Undergraduate Research Fellowship (x2)

For "PrefGenML: Preference Guided Image Generation with Relative Attributes"

2018

Pittsburgh Science and Technology Academy Valedictorian

Top of my high school class

Publications

Selected: Latest & Greatest

ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, Polo Chau
International Conference on Machine Learning (ICML). Vancouver, Canada, 2025.

[Project](#) [Demo](#) [PDF](#) [Code \(322 ⭐\)](#) [BibTeX](#) [🏆 Oral \(Top 1% of Submissions\)](#)

ManimML: Communicating Machine Learning Architectures with Animation

Alec Helbling, Duen Horng (Polo) Chau

IEEE VIS: Visualization Conference (IEEE VIS). Melbourne, Australia, 2023.

[Project](#) [PDF](#) [Code \(3108 ⭐\)](#) [BibTeX](#) [🏆 IEEE VIS 2023 Best Poster Award](#)

Diffusion Explorer: Interactive Exploration of Diffusion Models

Alec Helbling, Duen Horng (Polo) Chau

IEEE VIS: Visualization Conference (IEEE VIS). Vienna, Austria, 2025.

[Project](#) [PDF](#) [Code \(811 ⭐\)](#) [BibTeX](#)

All Publications

Transformer Explainer: Interactive Learning of Text-Generative Models

Aeree Cho, Grace C. Kim, Alexander Karpekov, Alec Helbling, Zijie J. Wang, Seongmin Lee, Benjamin Hoover, Duen Horng (Polo) Chau

IEEE VIS: Visualization Conference (IEEE VIS). Tampa Bay, USA, 2024.

[Project](#) [Demo](#) [PDF](#) [Code \(5268 ⭐\)](#) [BibTeX](#) [🏆 IEEE VIS 2024 Best Poster Award](#)

RenderBender: A Survey on Adversarial Attacks Using Differentiable Rendering

Matthew Hull, Haoran Wang, Matthew Lau, Alec Helbling, Mansi Phute, Chao Zhang, Zsolt Kira, Willian Lunardi, Martin Andreoni, Wenke Lee, Polo Chau

IJCAI 2025, Survey Track (IJCAI). 2025.

[Project](#) [PDF](#) [BibTeX](#)

Visualizing Convolutional Neural Network Protein-Ligand Scoring

Joshua Hochuli, Alec Helbling, Tamar Skaist, Matthew Ragoza, David Ryan Koes

Journal of Molecular Graphics and Modeling (JMGM). 2018.

[Project](#) [PDF](#) [BibTeX](#)

ClickDiffusion: Harnessing LLMs for Interactive Precise Image Editing

Alec Helbling, Seongmin Lee, Duen Horng (Polo) Chau

CVPR, Workshop on AI for Content Creation, 2024 (CVPR, Workshop). 2024.

[Project](#) [BibTeX](#)

ObjectComposer: Consistent Generation of Multiple Objects Without Fine-tuning

Alec Helbling, Evan Montoya, Duen Horng (Polo) Chau

NeurIPS 2023, Workshop on Machine Learning for Creativity and Design (NeurIPS 2023 MLCD Workshop). 2023.

[Project](#) [PDF](#) [BibTeX](#)

LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked

Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, Duen Horng (Polo) Chau

ICLR Tiny Paper Track (ICLR Tiny Paper). 2024.

[Project](#) [PDF](#) [BibTeX](#)

Manifold Contrastive Learning with Variational Lie Group Operators

Kion Fallah, Alec Helbling, Kyle A. Johnsen, Christopher J. Rozell

Transactions on Machine Learning Research (TMLR). 2023.

[Project](#) [PDF](#) [BibTeX](#)

PrefGen: Preference Guided Image Generation with Relative Attributes

Alec Helbling, Christopher John Rozell, Matthew O'Shaughnessy, Kion Fallah

Arxiv (Arxiv). 2023.

[Project](#) [PDF](#) [BibTeX](#)

Mobile Fitting Room: On-device Virtual Try-on via Diffusion Models

Justin Blalock, David Munechika, Harsha Karanth, Alec Helbling, Pratham Mehta, Seongmin Lee, Duen Horng (Polo) Chau

CVPR 2024, Workshop on Virtual Try-on (CVPR Workshop). 2024.

[Project](#) [BibTeX](#) [Invited Talk](#)

Oracle Guided Image Synthesis with Relative Queries

Alec Helbling, Christopher John Rozell, Matthew O'Shaughnessy, Kion Fallah

ICLR Workshop on Deep Generative Models for Highly Structured Data (ICLR DGMHSD). 2022.

[Project](#) [PDF](#) [BibTeX](#)

Mentoring

Fall 2023

Evan Montoya at Georgia Tech

BS in Computer Science, Georgia Institute of Technology

Methods for enabling compositionality of text-to-image generation systems.

Spring 2024

Aeree Cho, Grace Kim, Alexander Karpekov at Georgia Tech

MS in CS, BS in CS, MS in CS, Georgia Tech

Transformer Explainer: Interactive Learning of Text-Generative models

Grants and Funding

2023 — 2028

Human-in-the-loop Systems for Guiding Generative Models

NSF GRFP: National Science Foundation Graduate Research Fellowship Program

Co-PIs: Duen Horng (Polo) Chau

Funded \$80,000/year for 3 years (over of a 5 year period)

2021

Controlling Deep Generative Models with Pairwise Comparisons

President's Undergraduate Research Fellowship

Co-PIs: Chris Rozell

Funded \$1500/semester for 2 semesters

References

Dr. Polo Chau, Associate Professor

School of Computational Science and Engineering

Georgia Institute of Technology

cc.gatech.edu/~dchau/

Dr. Chris Rozell, Professor

School of Electrical & Computer Engineering

Georgia Institute of Technology
siplab.gatech.edu/rozell.html

Dr. Lukas Mandrake, Senior Researcher and Group Supervisor
Machine Learning and Instrument Autonomy Group
NASA Jet Propulsion Lab
ml.jpl.nasa.gov/people/mandrake/mandrake.html

Dr. Achille Fokoue, Distinguished Research Staff Member
Foundations of AI Reasoning Group
IBM Research
researcher.watson.ibm.com/researcher/view.php?person=us-achille