Ethan Shen

ethans03@cs.washington.edu | ethanlshen | ethanlshen

Education

University of Washington Seattle, WA

B.S. in Computer Science, B.A. in Mathematics, Minor in History Aug. 2022 - Present

- **GPA**: 3.99 / 4.00
- **Coursework**: Reinforcement Learning (Grad), Deep Learning (Grad), Natural Language Processing (Grad), Operating Systems, Data Structures and Parallelism, Databases, Software Design, Probability I and II, Linear Algebra, Multivariate Calculus
- **Activities**: AI Research Assistant, Teaching Assistant, Student Interviewer for CSE Faculty Hiring, CSE Student Advisory Council, Lavin Entrepreneurship Program

Skills

Languages	Java, Python, SQL, JavaScript, C, C++, HTML/CSS
Frameworks	PyTorch, Jax, HuggingFace, React, Node.js, NextJS, Flask, JUnit, scikit-learn
Dev Tools	Git, Google Cloud Platform, AWS, Azure, MongoDB, DynamoDB, Docker, Linux

Publications

Perception Tokens Enhance Visual Reasoning in Multimodal Language Models
Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, **Ethan Shen**, Dongping Chen, Linda Shapiro, Ranjay Krishna.
Under Review.

Superposed Decoding: Multiple Generations from a Single Autoregressive Inference Pass.
Ethan Shen, Alan Fan, Sarah M Pratt, Jae Sung Park, Matt Wallingford, Sham M Kakade, Ari Holtzman, Ranjay Krishna, Ali Farhadi, Aditya Kusupati.
NeurIPS 2024.

Are "Hierarchical" Visual Representations Hierarchical?
Ethan Shen, Ali Farhadi, Aditya Kusupati.
Workshop on Symmetry and Geometry in Neural Representations @ NeurIPS 2023.

Experience

RAIVN Lab @ UW Seattle, WA

AI Research Assistant (Prof. Ali Farhadi, Prof. Ranjay Krishna) Jun. 2023 - Present

- Working on additive compute strategies to accelerate RAG, image retrieval, and search.
- Developed a novel LLM decoding method using interpolated token embeddings to generate multiple outputs in a single inference pass, improving inference speed and accuracy for draft-based applications like Github Copilot.
- Created a suite of new hierarchical vision datasets and discovered that computer vision models can learn complex visual hierarchies without any special hyperbolic or adaptive training.

Amazon Seattle, WA

Software Engineer Intern Jun. 2024 - Sep. 2024

- Built a scalable GenAI pipeline with AWS Bedrock, Lambda, and DynamoDB to automatically process up to 15,000 financial documents a month, saving over 1200 hours of manual work per year.
- Created an internal invoice dataset of 45+ vendors and improved pipeline accuracy to 96% through prompt engineering, fine tuning, and heurestic filtering.

Papyrus San Francisco, CA

AI Engineer Intern Dec. 2023 - Mar. 2024

- Created a method to label speakers in transcripts using long-context knowledge from LLMs, with the feature becoming critical to the company's product.
- Built evaluation pipeline for transcription/translation using NextJS and Python and deployed it with AWS Lambda and Batch, saving 20 hours of testing monthly.

Sensor Systems Lab @ UW Seattle, WA

Robotics Research Assistant (Prof. Joshua Smith) Jun. 2022 - Jun. 2023

- Designed and built an acoustic levitator, a tool that uses ultrasonic sound to levitate fragile objects for scientific experiments.
- Programmed mathematical algorithms in Python and C++ to simulate, predict, and control the rotation of objects in the levitator, with a 15x speedup compared to existing simulations.

Mutorials Bellevue, WA

Software Developer Mar. 2020 - May 2022

- Helped found Mutorials, an ongoing science practice website with 3,400+ problems and 30,000+ user interactions.
- Implemented backend features like problem practice, client authentication, and user profiles with NodeJS and MongoDB.
- Designed and created frontend pages using HTML, EJS, Bootstrap, and Figma.

Honors & Awards

2023	**NeurIPS Travel Award**, Conference for Neural Information Processing Systems
2022	**FEEA Merit Scholarship**, Department of Veterans Affairs
2021	**42/45 Score (Top 8% Internationally)**, IB Diploma Programme
2021	**National Merit Finalist (Top 1% Nationally)**, National Merit Program

Professional Services

2024	**EMNLP 2024 Reviewer**
2024	**UW CSE: Student Interviewer for Faculty Hiring**
2023	**UW CSE: Teaching Assistant**
2022-2023	**UW CSE: Student Advisory Council Officer**

Presentations

Oct. 2024	**UW RAIVN Lab (Host: Ali Farhadi)**, Topic: Superposed Decoding
Aug. 2024	**AWS AI Labs CodeGen (Host: Zijian Wang)**, Topic: Superposed Decoding
Aug. 2024	**Amazon LLM Reasoning Group (Host: Linbo Liu)**, Topic: Superposed Decoding
Jul. 2024	**AWS AI Labs Quicksight (Host: Patrick Ng)**, Topic: Superposed Decoding