Zhengxu (Jason) Yan

Berkeley, CA 94608 | (509)330-6699 | jason.yan@berkeley.edu | Citizenship: U.S. Citizen | zhengxuyan.github.io

EDUCATION

University of California, Berkeley Expected Graduation: May 2025

Bachelor of Arts in Computer Science and Data Science GPA: 3.94/4.00, Dean's List, Honors to Date

- *Coursework: Principles & Techniques of Data Science (Data C8/C100), Computer Programs (CS 61A), Data Structures (CS 61B), Machine Structures (CS 61C), Discrete Math and Probability (CS 70), Optimization Models (EECS 127), Probability for Data Science (Data C140), Computer Security (CS 161), Operating Systems (CS 162), Efficient Algorithm (CS 170), Computer Vision (CS 180), Artificial Intelligence (CS 188), Machine Learning (CS 189), LLM Agents (CS 194-196).*

PUBLICATIONS

- **Yan, Z.**, Dube, V., Heselton, J., Johnson, K., Yan, C., Jones, V., Blaskewicz Boron, J., & Shade, M. (2024). Understanding older people's voice interactions with smart voice assistants: a new modified rule-based natural language processing model with human input. *Frontiers in digital health*, *6*, 1329910. https://doi.org/10.3389/fdgth.2024.1329910
- Jones, V. K., Yan, C., Shade, M. Y., Boron, J. B., **Yan, Z.**, Heselton, H. J., Johnson, K., & Dube, V. (2024). Reducing Loneliness and Improving Social Support among Older Adults through Different Modalities of Personal Voice Assistants. *Geriatrics (Basel, Switzerland)*, *9*(2), 22. https://doi.org/10.3390/geriatrics9020022

WORKING PAPERS

DiT-Serving: Infinitely Long Video Generation Engine with Brick Attention Enhancing Throughput via Continuous Batching

- Integrated Brick Attention into Diffusion Transformer (DiT) models to enable near-infinite video generation, supported by continuous batching to achieve increased throughput.
- Under the guidance of PhD candidate Michael Luo, advised by Dr. Ion Stoica, Prof. of Computer Science, I led the integration of Brick Attention with DiT models, enhancing the system's capacity for extended video generation, and implemented continuous batching techniques that significantly improved system throughput. The manuscript is in preparation for review.

Adaptive Operations Management in Buildings: A Reinforcement Learning Approach for Operational Adaptability in Healthcare Facilities

- Introduced a Reinforcement Learning-based framework to enable adaptive and integrated operations management in healthcare facilities, optimizing spatial, social, and operational performance through coordinated resource sharing.
- Supervised by Dr. Yehuda Kalay and Dr. Davide Schaumann, I utilized deep reinforcement learning (RL) and simulation to develop a smart building management system for the Cardiac Catheterization Lab at St. Bernardine Medical Center, significantly enhancing facility adaptability and operational efficiency. The manuscript is in preparation for journal review.

WORK EXPERIENCE

Teaching Assistant for Large Language Model Agents (CS 194-196) (Supervised by Dr. Dawn Song) August 2024 - Present

Department of Electrical Engineering and Computer Sciences at UC Berkeley *Berkeley, CA*

- Coordinated course logistics, managed lecture recordings, enhanced assignment materials, and provided academic support by addressing student inquiries.
- Assisted in organizing and coordinating the LLM Agents Hackathon, hosted by Berkeley RDI in conjunction with the LLM Agents MOOC, designed to foster innovation, expand the AI agent community, and advance LLM agent technology.

Research Assistant (Supervised by PhD candidate Michael Luo, advised by Dr. Ion Stoica) February 2024 - Present

UC Berkeley Sky Computing Lab *Berkeley, CA*

- Competitively selected through UC Berkeley EECS Diversifying Access to Research in Engineering (DARE) and Sky Summer Undergraduate Programs to join the DiT-Serving Project as a research assistant.
- Led a team of 5 research assistants in integrating Ring Attention and Brick Attention into Diffusion Transformer (DiT) models, pioneering scalable video generation techniques.
- Directed continuous batching strategies to optimize system throughput, enhancing performance efficiency across video processing requests.

Research Assistant (Supervised by Dr. Yehuda Kalay and Dr. David Schaumann) February 2024 - Present

UC Berkeley College of Environmental Design *Berkeley, CA*

- Competitively selected to join the Smart Hospital Project through UC Berkeley Undergraduate Research Apprentice Program.
- Led a team of 3 research assistants in the development and implementation of a deep reinforcement learning-based smart building management system for the Cardiac Catheterization Lab at St. Bernardine Medical Center.

Machine Learning (ML) Engineer Lead (Supervised by Dr. Stefano Bertozzi) January 2024 - Present

UC Berkeley School of Public Health *Berkeley, CA*

- Selected as lead computer scientist for Rapid Reviews: Infectious Diseases via the UC Berkeley CDSS Discovery Program.
- Independently leveraged a fine-tuned Large Language Model (LLM) to efficiently categorize and identify preprints within the RR\ID domain, significantly enhancing peer review efficiency.
- Led a team of 3 UC Berkeley CS students in utilizing LLM APIs to analyze and provide insights on medical preprints, supporting the peer review and decision-making process.
- Directed a team of 5 UC Berkeley CS students in engineering scripts to automate the collection of medical preprints from multiple servers, streamlining data acquisition and analysis workflows.
- Collaborated with the Dean of the UC Berkeley School of Public Health and researchers from UCSF to develop automated systems for the academic review of infectious disease papers, set to be published by MIT Press.

Research Assistant (Supervised by Dr. Valerie Jones) April 2023 – September 2023

University of Nebraska *Lincoln, NE*

- Collaborated with gerontology and nursing researchers to develop a natural language processing (NLP) model designed to automate the processing of speech-to-text data from user interactions with AI-enabled smart voice assistants.

Backend Engineer Lead October 2022 – December 2022

Coffee Tea, Inc. *Berkeley, CA*

- Led a team of 3 UC Berkeley computer science students in backend development for a social platform aimed at broadening access to the college application process by facilitating coffee chat connections between applicants and current college students.
- Directed the design and development of a comprehensive suite of backend REST APIs using FastAPI, Poetry, Alembic, and PostgreSQL, covering modules such as authentication, profile management, recommendation systems, and payment processing.

PROJECTS

Machine Learning

Multi-Agent LLM Trading System | *Python, Pytorch, Tensorflow, AutoGen* August 2024 – Present

- Developed a multi-agent trading system integrating LLMs with specialized agents for multimodal data processing, layered memory, and Retrieval-Augmented Generation, aiming to exceed standard prediction models by 5-7% in market accuracy.

IM2SPAIN Project | *Python* January 2024 – May 2024

- Employed nearest neighbors (k-NN) to predict geographic coordinates based on CLIP embeddings of geo-tagged images from Flickr, capturing diverse landscape features across Spain.

MNIST Competition | *Python, Pytorch* January 2024 – May 2024

- Engaged in the Kaggle MNIST classification challenge, leveraging a diverse array of machine learning techniques including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression, Multi-layer Perceptron (MLP), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN) to optimize prediction accuracy.

Language Identification RNN | *Python, Tensorflow* August 2023 – December 2023

- Engineered an RNN to identify word languages, dynamically processing variable-length character inputs to achieve over 81% test set accuracy.

DeFi Cryptocurrency Trends Analysis | *Python, Tensorflow* August 2023 – December 2023

- Developed an advanced DeFi cryptocurrency analysis tool with a Django-powered backend and React-based frontend, integrating LSTM neural networks for accurate time series forecasting of market trends.

Graph Partitioning for Unsupervised Learning | *Golang, AWS, Google Cloud* August 2022 – December 2022

- Developed graph partitioning algorithms to perform unsupervised machine learning tasks, using the Kernighan-Lin algorithm.
- Implemented a relational database and Golang backend on AWS and Google Cloud platforms to manage outputs and facilitate the running of algorithms.

Systems Programming and Software Development

Pintos Operating System | *C, Rust* January 2023 – May 2023

- Led the development of a comprehensive Pintos operating system, focusing on systems programming, memory allocation, resource management, file systems, networking, and security, enhancing system functionality and efficiency.

Secure Client Application | *Golang* January 2023 – May 2023

- Developed a secure client Golang application incorporating cryptographic primitives to manage authentication, file management, sharing, and access revocation, significantly enhancing data security and user privacy.

Gitlet Version Control System | *Java* January 2021 – May 2022

- Created “Gitlet,” a Git-like Version Control System, to streamline tracking and management of code changes across projects.

TECHNICAL SKILLS

Areas: Machine Learning, Deep Reinforcement Learning, Operating Systems, Natural Language Processing, Computer Vision
Languages: Python, Java, C/C++, Golang, SQL (Postgres), JavaScript, Rust, Bash, HTML/CSS
Developer Tools: Git, AutoGen, Stable Baselines, Docker, FastAPI, Django, Google Cloud Platform, AWS, Vercel
Libraries: TensorFlow, PyTorch, Pandas, NumPy, Matplotlib, Seaborn, CV2, Scrapy, OpenAI Gym