Junwei Yu

470-896-7237 | yujunwei04@berkeley.edu | linkedin.com/in/yujunwei1018 | yujunwei04.github.io/

Education

University of California, Berkeley — Berkeley, CA
B.A. in Computer Science, B.A. in Applied Mathematics — *Aug. 2022 - Dec. 2025*

- **Overall GPA**: 3.95/4.0
- **Honors**: EECS Honors Program, Upsilon Pi Epsilon, Dean's List
- **Selected Coursework**: Graduate Computer Vision, Deep Learning, Machine Learning, Convex Optimization, Algorithms, Probability Theory, Statistical Inference, Abstract Linear Algebra, Abstract Algebra, Real Analysis, Complex Analysis, Numerical Analysis, Machine Structures, Data Structures, Discrete Math, Data Science

Publications & Preprints

1. **Junwei Yu**, Trevor Darrell, and XuDong Wang, "UnSAMv2: Self-Supervised Learning Enables Segment Anything at Any Granularity," *under review at Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. [Paper] [Project Page]

2. Xiaoyu Li*, Yingyu Liang*, Zhenmei Shi*, Zhao Song*, and **Junwei Yu***, "Fast John Ellipsoid Computation with Differential Privacy Optimization," in *Conference on Parsimony and Learning (CPAL)*, Oral, 2025. [Paper]

3. Xiaoyu Li*, Zhao Song*, and **Junwei Yu***, in "Quantum Speedups for Approximating the John Ellipsoid," in *Quantum Information Processing Conference (QIP)*, 2025. [Paper]

** indicates alphabetical author order or equal contribution.*

Experience

Berkeley Artificial Intelligence Research — Nov. 2024 – Present
Undergraduate researcher, advised by Dr. XuDong Wang, Prof. Trevor Darrell — *Berkeley, CA*

- Designed and implemented the first unsupervised segmentation algorithm that discovers object instances and their corresponding granularity scales in a hierarchy-based manner, extending the divide-and-conquer mask discovery strategy to treat granularity as a relative and continuous concept.
- Proposed a self-supervised granularity training strategy that fine-tunes pretrained segmentation models such as SAM 2 to understand object granularity using only 0.02% extra parameters and 6,000 unlabeled images.
- Achieved state-of-the-art interactive and whole-image segmentation, surpassing SAM 2 by 26.0% and 37.7%, respectively, and unlocking granularity-controllable "segment anything" capabilities for downstream tasks.
- Explored research topics including self-supervised learning, representation learning, and vision-language models.

Simons Institute for the Theory of Computing — Apr. 2024 – Oct. 2024
Undergraduate researcher, advised by Dr. Zhenmei Shi, Dr. Zhao Song, Prof. Yingyu Liang — *Berkeley, CA*

- Researched and presented the first quantum algorithm that computes the John Ellipsoid. Our algorithm runs in $O(\sqrt{n}d^{1.5} + d^{\omega})$ time. In the tall dense matrix regime, our algorithm achieves quadratic speedup, resulting in a sublinear running time and significantly outperforming the current best classical algorithms.
- Proposed the first algorithm for fast John Ellipsoid computation that ensures differential privacy. Our work demonstrates the algorithm's convergence and privacy properties mathematically through a sequential privacy mechanism, providing a robust approach for balancing utility and privacy in John Ellipsoid computation.
- Explored theory research topics including concentration inequalities, dynamic attention, and neural tangent kernel.

Amazon — May 2025 – Aug. 2025
Software Development Engineer Intern — *Seattle, WA*

- Designed and developed a self-service internal dashboard from scratch individually to manage Amazon co-branded financial account applications across the world. It achieves queries and updates of clients' account-level financial data in real time, replacing manual ticket workflows and cutting average request turnaround by 80%.
- Architected an end-to-end serverless stack with AWS Lambda, IAM, SNS topic, API Gateway, DynamoDB, and CloudWatch; design scales automatically to >10K daily requests while keeping latency under 200 ms.
- Onboarded financial service APIs to Amazon payment internal MCP server that enables Claude to query APIs directly in the dashboard with appropriate authentication, boosting self-serve efficiency by 25%.

Teaching Experience

DATA C140: Probability Theory for Data Science Aug. 2024 – Dec. 2024
Teaching Assistant, supervised by Prof. Ani Adhikari *Berkeley, CA*

- Hold in-person office hour every week to answer students' questions on course contents and debugging.
- Taught concepts such as expectation, variance, Markov chain, tail bounds, central limit theorem, gamma distribution, moment generating function, maximum likelihood estimation, multivariate Gaussian, and regression.

Projects

Autoregressive Diffusion with Flow Matching | *Pytorch, CUDA* | [Project Page] Spring 2025

- Designed a compact, end-to-end framework that combines autoregressive sampling with flow-matching training to generate text-conditioned videos directly from Gaussian noise.
- Trained a model on the QuickDraw! dataset to predict the next stroke chunk from past chunks, enabling sketch video generation from text prompts. Leveraged a VAE latent space to improve training efficiency.

Convolution Code Optimization | *C, OpenMP, OpenMPI* Fall 2023

- Applied SIMD, OpenMP, Multithreading, and loop unrolling techniques to speed up convolution by 9.73 times.
- Implemented Gaussian blur algorithm with the fast parallel-version convolutions and applied it to video processing practices such as video blurring and video frames sharpening.

Technical Skills

Languages: Java, Python, PyTorch, C/C++, SQL, JavaScript, HTML/CSS, MATLAB, RISC-V
Frameworks: React, Node.js, Django, JUnit, AWS toolkits
Libraries: Tramsformers, Timm, Pandas, NumPy, Matplotlib, Seaborn, Sklearns