Ryan Tabrizi

(650) 804-6468 rtabrizi@berkeley.edu linkedin.com/in/rtabrizi ryantabrizi.com

Education

University of California, Berkeley **May 2025**

B.S. in Electrical Engineering and Computer Science *GPA: 3.80*

- **Graduate Coursework:** Vision-Language Models (In Progress), Computer Vision (A), NLP (A), Seminar: From Research to Startups (A)
- **Undergraduate Coursework:** Intro to ML (A+), CV and Computational Photography (A), Convex Optimization (A), Operating Systems (A), Computer Architecture (A), Data Structures (A), Electronic System Design (A), Linear Algebra and Differential Equations (A)
- **Distinctions and Leadership:** EECS Honors Program (Neuroscience Concentration), Eta Kappa Nu Honor Society, Dean's Honors List (4×), Berkeley AI Research, Launchpad Creative ML (Head of Education), Computer Science Mentors

Publications

An LLM Compiler for Parallel Function Calling [paper][code] **ICML 2024**

Sehoon Kim*, Suhong Moon*, **Ryan Tabrizi**, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, Amir Gholami

LLMCompiler: An open-source framework for efficiently executing multiple tasks in parallel. LLMCompiler manages dependencies and executes user-provided APIs, achieving up to 1.3x speedup over OpenAI's Parallel Function Calling, as well as up to 9% higher accuracy compared to ReAct.

TinyAgent: Function Calling at the Edge [paper][code] **EMNLP Demo 2024**

Lutfi Erdogan*, Nicholas Lee*, Siddharth Jha*, Sehoon Kim, **Ryan Tabrizi**, . . . , Kurt Keutzer, Amir Gholami

On-device deployment of fine-tuned small language models (<7B) to execute everyday user tasks on MacOS with LLMCompiler backend.

Experience

Berkeley AI Research - Computer Vision Group **May 2024 - Present**

Researcher *Berkeley, CA*

- Using REINFORCE and PPO for hyperparameter optimization advised by Prof. Angjoo Kanazawa; trained actor-critic network to automatically predict optimal Gaussian Splatting parameters for image fitting, eliminating manual tuning for each image.
- Led refactoring of Nerfstudio gsplat library, defining common API for different Gaussian Splatting methods to simplify splatting pipeline.

Berkeley AI Research - PALLAS Group **Sep 2023 - Sep 2024**

Researcher *Berkeley, CA*

- Built LLMCompiler (ICML 2024, 1.5K+ GitHub stars), an open-source parallel function calling framework for LLMs, with Prof. Kurt Keutzer.
- Achieved SOTA success rate on WebShop benchmark using LLMCompiler to call web agents in parallel for efficient environment exploration.
- Developed TinyAgent (EMNLP 2024 Demo), utilizing LLMCompiler for application-specific function-calling on small language models (<7B).

UC Berkeley EECS **Aug 2024 - Present**

Teacher Assistant *Berkeley, CA*

- Paid 10-hour TA for CS 180/280A "Computer Vision and Computational Photography," working alongside Prof. Alyosha Efros.
- Designed diffusion visual illusions project with DDPM and rectified flow in collaboration with UMich Prof. Andrew Owens.
- Conduct weekly office hours and project parties, and grade projects and exams for 300+ students as part of 8-person course staff.
- Incoming 10-hour TA for CS 280 "Graduate Computer Vision" with Prof. Angjoo Kanazawa and Prof. Jitendra Malik.

Insitro - Core Imaging Team **May 2023 - Aug 2023**

Research Scientist Intern *San Francisco, CA*

- Developed IM2MEA, a system that aligns 6k+ microscopy images with 2.5k+ microelectrode array (MEA) recordings to retrieve activation potential signals for target validation of neurodegenerative diseases.
- Achieved cross-modal mapping by utilizing META AI's DINO segmentation model on a dataset of 5k image-MEA data points to impute MEA aggregate metrics and image embedding centroids.
- Trained a transformer encoder-decoder on 2k MEA samples to reconstruct action potentials for downstream signal synthesis by encoding electrode-level and network-level MEA features.

NASA JPL - ML and Instrument Autonomy Team **May 2022 - Sep 2022**

Research Scientist Intern *Palo Alto, CA*

- Headed research and trade study of 5 image compression algorithms, including the ICER Progressive Wavelet Image Compressor and JPEG 2000, with 500+ preprocessed images across Mars, Miranda, and Ceres mission data.
- Designed compression algorithm analysis pipeline with Python wrappers for C codebases used in deep space exploration, utilizing structural similarity indexes and heat maps as benchmark metrics.
- Enabled and streamlined optimization of compression algorithms through pipeline's analytics for future missions within JPL's Machine Learning and Instrument Autonomy Group and NASA at large.

Projects

VeggieWorld | *Nerfstudio, gsplat, Blender* **Apr 2024**

- 'Veggified' 3D Gaussian Splatting scene reconstructions using vegetable 3D asset library for 3D style transfer.
- Compared CLIP ViT and human performance to asses veggified scene classification under different vegetable sizes and distribution densities.

Phodexr | *CLIP, BERT, ResNet, Hugging Face, WandB* **Sep 2021 - Dec 2021**

- Recreated Apple's camera roll semantic search with OpenAI's CLIP for lightweight album indexing from user text queries.
- Pre-trained a DistilBERT tokenizer and ResNet50 on COCO Captions dataset (300K+ image-caption pairs) for image retrieval.

Skills

Programming Languages: Python, Java, C, LaTeX, Scheme (LISP), RISC-V
Libraries/Frameworks: PyTorch, NumPy, Hugging Face, Nerfstudio, GSplat, LangChain, TensorFlow, Keras, scikit-learn, Pandas
Developer Tools: Git, AWS, Linux, WandB, Streamlit, Gradio, Isaac Gym, STM32Cube
Spoken Languages: English (native), Persian (B1), Spanish (B1)