# Yatong Bai

**Fresh Ph.D. in Generative AI (Diffusion), Robust Deep Learning, and Optimization**

**yatong_bai@berkeley.edu**
Website: bai-yt.github.io
LinkedIn: linkedin.com/in/yatong-bai

## About

I am a machine learning researcher who recently obtained my Ph.D.
I work in two main directions, spanning modalities including image (CV), language (NLP), and audio/music.
- **Generative (diffusion) models.** I accelerate them and align them with human preferences.
- **Adversarial robustness.** I investigate model vulnerabilities and seek ways to defend against them without losing performance.

## Education

**University of California, Berkeley,** Doctor of Philosophy & Master of Science — *Aug 2020 – May 2025*
- **Advisor:** Somayeh Sojoudi. **GPA:** 4.00 / 4.00
- **Areas:** Computer science, deep learning, diffusion models, generative AI (audio/music/vision), adversarial robustness, optimization, control.
- **Dissertation:** *Efficient and Reliable Optimization for Deep Learning and Media Generation* (bai-yt.github.io/files/dissertation.pdf).

**Georgia Institute of Technology,** Bachelor of Science — *Aug 2016 – Aug 2020*
- **Double major** in Computer Engineering and Mechanical Engineering. **GPA:** 4.00 / 4.00

## Work Experience

**ByteDance,** Research Scientist — *San Jose, CA, July 2025 – Current*
- Working on music foundation models.

## Internships

**Adobe Research,** Intern Research Scientist — *San Francisco, CA, May 2024 – Jan 2025*
- Developed DRAGON, an on-policy reinforcement learning framework for AI media/music generation models compatible with reward functions that evaluate either individual generations or generated distributions. Paper under submission (arxiv.org/abs/2504.15217).
- Leveraged this framework to design novel reward functions that steer generations toward "exemplar sets".
- Collected a music preference dataset with a Slack App and modeled human preference to finetune/evaluate music generation models.
- Improved diffusion-based music generation to >60% win rate with sparse/zero human preference data or additional high-quality music.

**Microsoft** Applied Science Group, Research Intern — *Redmond, WA, May 2023 – Aug 2023*
- Accelerated diffusion-model-based text-to-audio generation 400x with minimal performance drop via "consistency distillation".
- Innovatively combined the consistency distillation framework and "classifier-free guidance" to distill high-performance models.
- The distilled model needs only one neural net query (diffusion models need 400), enabling improving audio semantics by optimizing audio-space losses. We use the CLAP score loss as an example and use objective and subjective evaluation to show its effectiveness.
- Published as "ConsistencyTTA" at INTERSPEECH 2024 (arxiv.org/abs/2309.10740).
- Code open-sourced at github.com/Bai-YT/ConsistencyTTA. Model checkpoints at huggingface.co/Bai-YT/ConsistencyTTA.
- Live demo at huggingface.co/spaces/Bai-YT/ConsistencyTTA. Project website and example generations at consistency-tta.github.io.

**Scale AI,** Machine Learning Research Intern — *San Francisco, CA, May 2022 – Dec 2022*
- Researched on constructing an e-commerce dataset with 15 million image-caption pairs and processed captions with language models.
- Evaluated the dataset and provided benchmark results with supervised and self-supervised image classification, object detection, image reconstruction, and generation methods (in PyTorch); visualized the embedding space with dimension reduction methods.
- Used these results to characterize the distribution shift of our data from existing datasets. Preprint paper arxiv.org/abs/2401.04575.

**Honda Aircraft Company,** Engineering Intern — *Greensboro, NC, May 2019 – Aug 2019*
- Flap linkage dynamic simulations in MSC ADAMS; Stress/deflection/kinematics evaluations in CATIA via Finite Element Analyses (FEA).
- Defined the flap skew and asymmetry warning thresholds and designed a flap control logic in MATLAB.

**Tesla, Inc.,** Engineering Intern — *Palo Alto, CA, May 2018 – Aug 2018*
- Implemented scripts that convert simulation models between different tolerance stack-up (GD&T) simulators.

## Publications

**DRAGON: Distributional Rewards Optimize Diffusion Generative Models**
**Yatong Bai**, J Casebeer, D Tran, S Sojoudi, and NJ Bryan. Preprint, 2025. arxiv.org/abs/2504.15217
- See the Adobe Research internship (listed above) for details.

### ConsistencyTTA: Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation

**Yatong Bai**, T Dang, D Tran, K Koishida, and S Sojoudi. In *INTERSPEECH*, 2024. arxiv.org/abs/2309.10740

- See the Microsoft internship (listed above) for details.

### Ranking Manipulation for Conversational Search Engines

S Pfrommer*, **Yatong Bai***, T Gautam, S Sojoudi (*equal contribution). arxiv.org/abs/2406.03589

In *Conference on Empirical Methods in Natural Language Processing (EMNLP **Oral**, top-10% accepted papers)*, 2024.

- Consider conversational search engines (CSE), powered by large language models (LLM) with retrieval augmented generation (RAG).
- When prompted to recommend websites, CSEs simultaneously pay attention to website topic, content, and location in context.
- By embedding spurious characters in HTML code, a website can deceive CSEs to boost its recommendation ranking in the response.
- Proposes RAGDOLL, an e-commerce website dataset for evaluation.
- Dataset open-sourced at huggingface.co/datasets/Bai-YT/RAGDOLL. Project code github.com/spfrommer/cse-ranking-manipulation.

### MixedNUTS: Training-Free Accuracy-Robustness Balance via Nonlinearly Mixed Classifiers

**Yatong Bai**, M Zhou, VM Patel, and S Sojoudi. In *Transactions on Machine Learning Research (TMLR)*, 2024. arxiv.org/abs/2402.02263

- Balances neural network classifiers' (un-attacked) clean accuracy and adversarial robustness without additional training, reducing error rate by up to 31% compared to a standalone model.
- Achieved by proposing "nonlinear base model logit transformations" for "mixed classifiers (introduced in the two entries below)".
- The transformations augment the robust base classifier's benign confidence property, thereby balancing accuracy and robustness.

### Improving the Accuracy-Robustness Trade-Off of Classifiers via Local Adaptive Smoothing

**Yatong Bai**, BG Anderson, A Kim, and S Sojoudi. In *SIAM Journal on Mathematics of Data Science (SIMODS)*, 2024. arxiv.org/abs/2301.12554

- Based on the work below, we further introduce a "mixing network" to adjust the mixing strengths for benign and attacked inputs differently, further improving the accuracy-robustness trade-off. As of submission, on CIFAR-10 and CIFAR-100 datasets, adaptive smoothing is the second most robust method on RobustBench, with noticeably higher clean accuracy than all other works.
- Project code open-sourced at github.com/Bai-YT/AdaptiveSmoothing.

### Mixing Classifiers to Alleviate the Accuracy-Robustness Trade-Off

**Yatong Bai**, BG Anderson, and S Sojoudi. In *Learning for Dynamics & Control Conference (L4DC)*, 2024. arxiv.org/abs/2311.15165

- By building a "mixed classifier", specifically mixing the output probabilities of an accurate (often non-robust) classifier and a robust classifier, we greatly alleviate the accuracy-(adversarial) robustness trade-off and achieve certified robustness.
- Our analysis shows that the robust base classifier (RBC)'s prediction confidence is the main source of this improvement. Specifically, the RBC is more confident in correct examples than incorrect ones. The mixed classifier thus puts more trust in correct predictions.

### Efficient Global Optimization of Two-Layer ReLU Networks: Adversarial Training and Quadratic-Time Algorithms

**Yatong Bai**, T Gautam, and S Sojoudi. In *SIAM Journal on Mathematics of Data Science (SIMODS)*, 2022. arxiv.org/abs/2201.01965

- 2021 INFORMS Data Mining Best Paper Competition (Student Track) Runner-up (2nd out of 48 papers).
- We develop ADMM algorithms for the "convex training" formulation to efficiently train one-hidden-layer neural networks with global optimality via convex optimization. We prove that the proposed algorithms polynomially improve the computational complexity.

### Initial State Interventions for Deconfounded Imitation Learning

S Pfrommer**, Yatong Bai**, H Lee, and S Sojoudi. In *IEEE Conference on Decision and Control (CDC)*, 2023. arxiv.org/abs/2307.15980

- Imitation learning agents suffer from causal confusion. We use a beta-VAE neural net to obtain disentangled latent representations underlying the observations, and use a statistical test to mask confounding latent variables so that the agent performs significantly better when the observations are confounded.

### Practical Convex Formulation of Robust One-Hidden-Layer Neural Network Training

**Yatong Bai**, T Gautam, Y Gai, and S Sojoudi. In *American Control Conference (ACC)*, 2022. arxiv.org/abs/2105.12237

- We leverage the duality theory and robust optimization techniques to develop efficient convex optimization formulations that train robust one-hidden-layer ReLU neural networks via adversarial training.
- Our method demonstrates improved adversarial robustness on common datasets, including CIFAR-10.

### Let's Go Shopping (LGS) – Web-Scale Image-Text Dataset for Visual Concept Understanding

**Yatong Bai**, U Garg, A Shanker, H Zhang, S Parajuli, E Bas, I Filipovic, AN Chu, ED Fomitcheva, E Branson, A Kim, S Sojoudi, K Cho. *Preprint*, 2024. arxiv.org/abs/2401.04575

- See the Scale AI internship (listed above) for details.

## ACADEMIC ACTIVITIES

- 43 Paper Reviews: Journals – TMLR (2024), SIMODS (2025)
  Conferences – CDC (2022, 2023), ICML (2023, 2024, 2025), CCTA (2023), NeurIPS (2023, 2024), ICLR (2024, 2025).
- Teaching (TA): IEOR 160: Nonlinear and Discrete Optimization – Spring 2022, Fall 2022, Fall 2023 (UC Berkeley).
  EECS 127: Optimization Models in Engineering – Fall 2024 (UC Berkeley).
- Talks/Posters: INTERSPEECH 2024, L4DC 2024, ACC 2022, CCTA 2023, INFORMS 2021, and MOPTA 2021 conferences,
  NorCal Control Workshop, Tsinghua-Berkeley Shenzhen Institute (TBSI) poster symposium, TBSI guest lecture.

## AWARDS

| | |
|---|---|
| Outstanding Graduate Student Instructor Award (UC Berkeley) | *March 2025* |
| The Henry Lurie Family Fellowship | *May 2024* |
| INFORMS Data Mining Best Paper Competition (student track) Runner-Up Prize | *Oct 2021* |
| UC Berkeley Graduate Division Block Grant Fellowship | *April 2021* |
| Georgia Tech School of ECE Roger P. Webb ECE Senior Scholar Awards | *April 2021* |

## UNDERGRADUATE EXPERIENCE

- I researched with the Meaud Research Group, the TINKER Group, and the GT Off-Road racing team during undergraduate studies.
- Compiled SPEC 2017 computer architecture benchmark into GCC-ARM binary programs and used Gem5 simulator (C++) to build debug traces.
- Built Graphical User Interfaces (GUIs) for a cochlear dynamics simulator in MATLAB. The GUIs controlled simulations, logged and processed experiment data, and visualized the simulation results.
- Implemented the avionics system of a novel unmanned "Monocopter" and a PID-controlled testbed using C++. The avionics filtered noisy magnetometer readings to accurately recover aircraft heading and control the actuators. Also developed a Windows C# GUI.

## CODING

I regularly code in Python (PyTorch) and MATLAB and write in LaTeX, often working on remote cloud computing machines.
Other languages I've used include HTML, C, C++, Java, and R.