John Robertson

Curriculum Vitae

johntrob14.github.io

john.robertson@utexas.edu

Education

Doctor of Philosophy in Electrical / Computer Engineering **2025-Present**
University of Texas at Austin *Advised by Dr. Haris Vikalo & Dr. Atlas Wang*
Honors: Charles W. and Margaret A. Tolbert Endowed Fellowship GPA: 3.9
Graduate Coursework:

- Applied Probability I, Statistical Machine Learning, Stochastic Processes I, Deep Learning, Deep Reinforcement Learning, and Generative Models.

Bachelor of Science in Electrical / Computer Engineering **2025**
University of Texas at Austin GPA: 3.6

Research Experience

Graduate Researcher 2025-Present
University of Texas at Austin *Advised by Dr. Haris Vikalo & Dr. Atlas Wang*

- Developing interpretable machine learning methods for AI safety, with additional interests in healthcare and computational biology.
- Currently spearheading multiple works in activation engineering and DNA modeling.

Undergraduate Researcher 2022-2025
University of Texas at Austin *Advised by Dr. Haris Vikalo*

- Developed transformer-based architectures (XVir, NextVir) to detect oncoviral DNA mutations in cancer samples, achieving state-of-the-art performance in viral DNA classification.
- Co-authored one publication (XVir) and first-authored a second (NextVir) in computational biology journals.

AI Research Intern 2024
Kilby Labs, Texas Instruments *Advised by Dr. Arthur Redfern*

- Developed two patent-pending works on efficient deep learning for edge devices as the sole undergraduate intern.
- Designed TIedNet, an architecture using shared weights and LoRA-like perturbations for memory-efficient image classification.
- Created Conditional PTQ, a method to improve post-training static quantization by predicting optimal scales for individual samples.

Publications and Manuscripts

When Is Rank-1 Steering Cheap? Geometry, Granularity, and Budgeted Search 2026

- *In submission. Preprint forthcoming on arXiv.*
- First author. Formalized rank-1 activation steering as a budget-constrained search over intervention layer and coefficient, and introduced *concept granularity* as a predictor of intervention difficulty.
- Proposed the GRACE workflow (Granularity- and Representation-Aware Concept Engineering) for cheaper, more reliable steering.
- Early version presented as a poster at Amazon's 2026 *Trusted AI Symposium.*

NextVir: Enabling Classification of Tumor-Causing Viruses with Genomic Foundation Models 2025

- *John T. Robertson, Shorya Consul, Haris Vikalo. PLOS Computational Biology, 2025.*
- First author. Led project adapting genomic foundational models with LoRA fine-tuning for viral mixture separation.

XVir: A Transformer-Based Architecture for Identifying Viral Reads from Cancer Samples 2025

- *Shorya Consul, John T. Robertson, Haris Vikalo. Journal of Computational Biology, 2025.*
- Presented at the CNB-MAC workshop, ACM Conference on Bioinformatics (2023).
- Co-authored work developing a transformer model for classifying oncoviral DNA.

Professional Experience

Course Assistant, Probability and Random Processes 2026
University of Texas at Austin *Instructor: Dr. Vivek Telang*

- TA for Probability and Random Processes in Shinjuku, Japan at J. F. Oberlin University.

Undergraduate Course Assistant 2024
Digital Signal Processing, University of Texas at Austin

- Assisted with teaching Digital Signal Processing material including Fourier Transforms, adaptive filters, and FFT.
- Responsible for grading homeworks and hosting office hours twice a week.

Artificial Intelligence Engineer Intern 2023
ClearBlade

- Deployed an IoT ML framework to monitor shoppers and notify service workers.
- Integrated Vertex-AI into ClearBlade's IoT core and developed an open-source alternative for ML training and deployment.

AI/ML Intern 2022
Jacinto-Edge AI, Texas Instruments

- Gained foundational machine learning experience on the Jacinto-Edge AI team.
- Studied real-time object detection on edge devices and built robotics demonstrations.

Technical Skills

Programming Languages: Python, C/C++, Assembly, Golang, Java, Typescript/JS, SQL

Software and Tools: Linux, Docker, Git, KubeFlow, GCP/Vertex-AI, Pandas, OpenCV, TensorFlow, PyTorch, Scikit-Learn, Agentic Programming

Projects

TWIIRL: Token-Wise Interpretable Interventions via Reinforcement Learning 2026

- Reformulated activation steering as a token-level decision problem; trained a small GRU controller via offline preference-based RL to emit per-token coefficients on a fixed diffmeans direction, under an explicit KL trust region.
- On Gemma 2 9B, strictly dominates fixed-coefficient Persona Vectors on the concept–coherence Pareto frontier at less than 0.01% per-token compute overhead.

DNA-ADLM: Anchored Diffusion for DNA Inpainting 2026

- Built a masked discrete diffusion language model (MDLM-style backbone, DiT denoiser) for reconstructing missing DNA spans given partial observations.

- Adapted anchored posterior sampling so observed anchor tokens stay pinned at exact equality while non-anchor positions are iteratively resampled, concentrating compute on uncertain regions.
- Pretrained on chromosome 11 with 5-mer tokenization; evaluated on inpainting tasks with 1–10 contiguous masked gaps.

Audio Spectrogram Transformer + MIL: Interpretable ALS Severity Classification 2025

- Won second place in the Speech Analysis for Neurodegenerative Diseases Grand Challenge.
- First author; accepted to IEEE ICASSP 2026 as an oral presentation (unpublished due to attendance conflicts).

FlareSight *(UT ECE Capstone Project)* 2024

- Led 5 engineers to develop an indoor multi-modal ML framework for sub-minute fire detection and localization.

DoxxCLIP 2023

- Led a team of 5 data scientists investigating ethical concerns of image geolocation models, developing a novel multi-armed bandits adversarial attack.