

# Vatsal Baherwani

New York, NY | [vatsal0.github.io](https://vatsal0.github.io) | [vatsalbaherwani@gmail.com](mailto:vatsalbaherwani@gmail.com) | [linkedin.com/in/vatsal-baherwani](https://linkedin.com/in/vatsal-baherwani) | [github.com/vatsal0](https://github.com/vatsal0)

## EDUCATION

---

### New York University

Sep 2025 – Present

PhD, Computer Science

Coursework: Bayesian Machine Learning, Linear Algebra

### University of Maryland

Aug 2021 – May 2025

B.S., Computer Science • Computational Finance Minor • Cumulative GPA: 3.98/4.0

Graduate Coursework: Foundations of Deep Learning, AI/ML at Scale

Teaching Assistant for BUFN400: Introduction to Financial Markets and Financial Datasets

## PUBLICATIONS

---

Video Diffusion Models Encode Motion in Early Timesteps, *Under Review at AAAI 2026*

**V. Baherwani**, Y. Ren, A. Shrivastava

DynaGuard: A Dynamic Guardrail Model With User-Defined Policies, *NeurIPS 2025*, *arXiv:2509.02563*

M. Hoover, **V. Baherwani**, N. Jain, K. Saifullah, J. Vincent, C. Jain, M.K. Rad, C.B. Bruss, A. Panda, T. Goldstein

Dense Backpropagation Improves Training for Sparse Mixture-of-Experts, *NeurIPS 2025*, *arXiv:2504.12463*

A. Panda\*, **V. Baherwani\***, Z. Sarwar, B. Thérien, S. Sahu, S. Rawls, S. Chakraborty, T. Goldstein

Racial and Gender Stereotypes Encoded Into CLIP Representations, *ICLR 2024 Tiny Papers Track*

**V. Baherwani**, J. Vincent

## RESEARCH EXPERIENCE

---

### Center for Human-Compatible AI – University of California, Berkeley

*Interpretable Hypothesis Generation for Image Data*

May 2025 – Aug 2025

Supervisor: Dr. Emma Pierson - Department of Electrical Engineering and Computer Science

- Implemented auto-interpretability pipeline and evaluation framework for image embedding models
- Trained sparse autoencoders to discover predictive features and generated interpretable hypotheses using VLMs
- Implemented web scraping and preprocessing tools for 6 data sources, resulting in >1 million total images

### University of Maryland

*Guardian Models for LLM Rule Compliance*

Aug 2024 – May 2025

Supervisor: Dr. Tom Goldstein - Department of Computer Science

- Created synthetic data pipeline to generate examples of LLM conversations following or violating specified rules
- Applied supervised fine-tuning on Qwen 4B and 8B models with >10,000 synthetic compliance examples
- Developed multi-GPU RL environment using GRPO and verifiable rewards to correctly detect rule violations
- Achieved F1 score of 81% in detecting unsafe LLM output and 73% in classifying compliance for arbitrary rules

*Stable Training for Sparse Mixture-of-Experts Language Models*

Jul 2024 – May 2025

Supervisor: Dr. Tom Goldstein - Department of Computer Science

- Proposed novel estimator for dense MoE gradient with 99% accuracy using moving average of sparse gradients
- Developed efficient Triton kernel implementation for gradient estimation, minimizing FLOPs overhead to <0.5%
- Trained 2B parameter LLMs with gradient estimator on 10B tokens of Fineweb dataset on Polaris supercomputer
- Improved training convergence speed by 9% and pretraining benchmark performance by 3% after 320B tokens

*Supervisor: Dr. Abhinav Shrivastava - Department of Computer Science*

- Traced motion learning in diffusion models by applying DDIM inversion and resampling at various timestep ranges
- Discovered and validated the property of video diffusion models encoding motion independently in early timesteps
- Developed new motion customization method by restricting diffusion model fine-tuning to only early timesteps
- Conducted ablation study to localize motion information in attention layers, reducing training parameters by >50%

*Understanding Historical Determinants of Market Liquidity and Depth*

*Sep 2023 – Dec 2023*

*Supervisor: Dr. Pete Kyle - Department of Finance*

- Developed Python library for loading and merging high-frequency trading data and calculating financial statistics
- Modeled stock market liquidity and market depth with time series regression on 1 terabyte of historical market data
- Analyzed time series autocorrelation, reduced noise with outlier filters, and implemented parallel data processing
- Aggregated monthly liquidity and depth for >5,000 companies, revealing a logarithmic correlation ( $r^2=0.95$ )

---

## INDUSTRY EXPERIENCE

### **Wolverine Trading**

*Software Engineer Intern*

*May 2023 – Aug 2023*

- Designed snapshot service to deliver daily historical price and market data for 50,000+ traded option contracts
- Improved average request speed by >75% and reduced database request load by >80% with server-side caching
- Built trader-facing client with .NET framework to query and cache market snapshots and listen to real-time updates
- Implemented client-server RPC communication and publisher/subscriber API for sending intraday market updates

### **Bloomberg**

*Software Engineer Intern*

*May 2022 – Aug 2022*

- Optimized calculation of 72 daily indicators for 3 million fixed-income securities with Apache Airflow workflows
- Parallelized execution and implemented autonomous error handling, reducing average calculation runtime by 35%

*Software Engineer Intern*

*Sep 2020 – Aug 2021*

- Developed Java Spring REST API to create, read, update, and delete 900+ Bloomberg Law account permissions
- Reduced service login time by >200ms through identifying and removing 300+ unused or expired permissions

---

## PERSONAL PROJECTS

### **UnityPack**

- Extracts assets including meshes, textures, and animations from compiled Unity games into glTF format JSON files
- Decompresses bit-packed mesh data, calculates animation transform matrices, and generates scene graphs

**NBA Shot Selection Analysis** – [vatsal0.github.io/cmssc320final/](https://vatsal0.github.io/cmssc320final/)

- Data cleaning, visualization, and regression analysis on 30 NBA teams predicting win rate with 82% accuracy
- SVM classification and k-means clustering to evaluate and rank shot efficiency from 16 areas on the NBA court

**MNIST Digit Classifier** – [github.com/vatsal0/mnist-digit-classifier](https://github.com/vatsal0/mnist-digit-classifier)

- Neural network written from scratch in C with implementations of regularization and mini batch gradient descent
- Classifies 10,000 handwritten digits from the MNIST image database test set with 97.1% accuracy

---

## ADDITIONAL INFORMATION

**Certifications:** Stanford Machine Learning, Kaggle Deep Learning, Bloomberg Market Concepts

**Personal Blog:** [medium.com/@vatsalbaherwani](https://medium.com/@vatsalbaherwani)