Vivek Mirani

vivek-mirani | vivekmirani19@gmail.com | (+91) 8017965515

Research Interests

Generalizable Robot Learning | Safe Autonomy | Efficient Inference

Education

Indian Institute of Technology, Kharagpur *Nov 21 - Jul 25*
Bachelor of Technology in Mechanical Engineering with Spl. in Artificial Intelligence and Applications CGPA: 8.60/10
Honors: Institute Order of Merit - Technology, O.P. Jindal Scholarship Nomination Kharagpur, India
Courses: Robotics, Graphical & Generative Models, Advanced Learning Paradigms for AI, NLP

Publications

- Garima Bansal*, **Vivek Mirani***, Arindam Biswas, Pabitra Mitra, Amaljith E V. *HEX: Merging Heavy-Hitters and Expanders for Adaptive KV Cache Optimization in Long-Context Inference.* Under Review at **ICLR 2026**.
- **Vivek Mirani***, Garima Bansal*, Arindam Biswas, Pabitra Mitra, Amaljith E V. *GEAR-X: Expanders for Next-Gen KV Cache Compression.* Accepted at **NeurIPS 2025 Efficient Reasoning Workshop**.

Experience

Research Intern | IISc Bangalore | Artificial Intelligence and Robotics Laboratory *May 23 - Present*
Control and Navigation for Autonomous Underwater Vehicles (Prof. Suresh Sundaram, AE) Bengaluru, India
– Developed a cascaded PID controller in Simulink by decoupling and linearizing state space equations about the trim state
– Implemented way-point navigation & static obstacle avoidance with Dynamic Window Approach using sonar feed in Gazebo
– Formulating modality-agnostic navigation policies by aligning sparse point clouds & dense visual data in shared latent space

Undergraduate Researcher | IIT Kharagpur | Bachelor Thesis *Aug 24 - Present*
Efficient LLM Training and Inference using Structured Sparsity (Prof. Pabitra Mitra, CSE) Kharagpur, India
– Investigated random and structured sparsity up to 85% on LoRA matrices to fine-tune RoBERTa with near-baseline accuracy
– Introduced expander-graph sparsity in KV cache, with spectral guarantees for stable CoT reasoning & long-context scenarios
– Complemented expanders with input-adaptive mask to store heavy & recent hitters in full precision, 3-bit quantizing the rest

Quantitative Strategist | Goldman Sachs | Global Markets *Jul 25 - Present*
Inventory Optimization for Synthetics Products Group Bengaluru, India
– Modeled the inventory management problem as a Mixed Integer Linear Program (MILP) to minimize cost & optimize holdings
– Linearized non-convex constraints using Big-M formulation with integer variables, reducing complexity and solver runtime

Projects

Conservative Implicit Q-Learning | Term Project *Mar 25 - Apr 25*
Supervisors: Prof. Plaban Kumar Bhowmick & Prof. Adway Mitra
– Merged IQL's in-support expectile updates & CQL's off-support Q-penalty, mitigating overestimation & distributional shift
– Implemented in PyTorch with expectile-based value fitting, advantage-weighted cloning & clamped penalty on OOD samples
– Evaluated on PyBullet D4RL benchmarks; achieved faster convergence, smoother returns compared to standalone methods

Approximate Inference in GMs | Term Project *Feb 25 - Mar 25*
Supervisors: Prof. Adway Mitra & Prof. Jiaul Houque Paik
– Designed a GNN-based inference model for MRFs, outperforming MLPs and nearing Gibbs accuracy with substantial speedup
– Achieved very low MSE with transformer attention, dynamic edge updates in message passing generalizing to 100-node graphs
– Optimized inference using label propagation, graph coarsening & spanning trees to scale beyond 30-node dense topologies

UAVs for Indoor Agriculture Challenge (ICUAS 2024) | Aerial Robotics Lab *Jan 24 - Feb 24*
Supervisor: Prof. Somesh Kumar
– Implemented waypoint tracking and obstacle avoidance in simulation, using a Finite State Machine for stable state transitions
– Optimized UAV trajectory using stochastic gradient descent, reducing travel time by 10% and energy consumption by 16%

Competitions

Inter IIT Tech Meet 13.0 | IIT Bombay | Captain *(Alphas on BTC & ETH Crypto Market)* *Oct 24 - Dec 24*
– Developed cross-asset strategies with 5+ Sharpe using filter crossovers (Biquad, LSMA, Gaussian) and technical indicators
– Achieved sub-15% max drawdown using dynamic TP/SL for different market phases minimizing risk in adverse conditions

Inter IIT Tech Meet 12.0 | IIT Madras | Gold *(University Project Ecosystem Challenge)* *Oct 23 - Dec 23*
– Created questionnaire, applied LDA for Topic Modeling, Binomial Tests for validation, and devised app-ranking algorithm
– Executed Monte Carlo simulation, quantifying the strategy impact on business goals through analysis of KPI distributions

Formula Bharat 2023 *Jan 23 - Feb 23*
– Collaborated in a team of over 30+ students in the designing and manufacturing of a Formula Student Combustion Vehicle
– Secured 8th rank overall & 3rd place in the Cost and Manufacturing Event among 70+ Formula Student teams nationally

Skills

Languages: Python, C/C++, Java, AVR C **Software**: GitLab, Colab, W&B, SolidWorks, Eagle, ROS, MATLAB, Simulink
Frameworks: Numpy, Pandas, NLTK, PyTorch, TensorFlow, Scikit-Learn, Transformers, OpenAI, PEFT, Gensim, OpenCV

Volunteering

- Headed a 40+ team to promote Robotics & AI; mentored 200+ students in Autonomous Robotics, Computer Vision & ML
- Led the hardware team at Aerial Robotics Lab & conducted sessions on quadrotor building & microcontrollers for sophomores