




Raghav Goel

raghavsyz@gmail.com | (412)-519-8580 |   

EDUCATION

CARNEGIE MELLON UNIVERSITY

School of Computer Science, The Robotics Institute

Master of Science in Robotics (MSR) (GPA: 3.95/4.0)

Coursework: Optimal Control and **Reinforcement Learning**, Kinematics Dynamics and Controls, **Machine Learning (PhD)**, Advance Dynamics and Simulator Design, Convex Optimization, **Visual Learning** and Recognition

Pittsburgh, PA

Dec 2022

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI (IIITD)

Bachelor of Technology in Electronics & Communication (Dept. Rank 1/70) (CGPA: 9.18/10)

Relevant Coursework: Machine Learning in Real Time Control, Nonlinear Control in Robotics, Adaptive Control in Robotics, Linear Optimization, Computer Vision, Statistical Signal Processing, Dynamical Systems, VLSI Design Flow, Embedded Logic Design

New Delhi, India

May 2020




EXPERIENCE

QUALCOMM, AI RESEARCH

Senior Deep Learning Researcher

5575 Morehouse Drive San Diego, CA, 92121, USA

23 Jan 2023-Present



- **Efficient LLM System** (PI: Mingu Lee, Chris Lott):
 - **Injected On-the-fly Speculative Decoding** (*under submission ICML 2025*)
 - Developed a training-free plug-and-play inference acceleration algorithm for self-speculative decoding, outperforming existing methods like lookahead decoding.
 - **CAOTE: efficient KV-cache eviction using attention-output** (*under submission ICML 2025*)
 - Proposed an **optimization framework** for efficient KV-cache eviction using change in attention-output combining both key and value information, enhancing performance of token-eviction methods like H2O, TOVA, SnapKV.
 - **Multi-modal (Self)-Speculative Decoding** (*CVPR 2024 Oral @ eVLM workshop*) [8] 
 - Pioneered **speculative decoding for multi-modal language models**, achieving a theoretical speed-up of ~2.4x.
 - **Recursive Speculative Decoding** (*ICLR 2024 LLM-Agents workshop*) [10] 
 - Introduced tree-based speculative decoding with sampling while maintaining base model output distribution.
 - **Direct Alignment of Draft model for Speculative Decoding using new loss** (*ICLR 2024 ME-FOMO workshop*) [9] 
 - Combined policy gradient with total-variation distance (TVD) to propose new distillation training loss TVD++, outperforming KLD, TVD based losses by up to 30%.
- **Compiler Optimization team** (PI: Will Zeng, Chris Lott):
 - Reduced DDR memory usage for deep-learning models run on memory-bound devices using an integer optimization problem
 - Improving scheduling of AI models on memory-bound devices through novel constraint reinforcement-learning algorithm.
 - **Accomplishments:** Proposed a new method for solving Schedule-aware Pinning (SAP) problem with near optimal performance without using licensed solvers with speed comparable to licensed solvers.

BIROBOTICS LAB, CMU (PI: Professor [Howie Choset](#))

Research Assistant

Pittsburgh, PA

Jan 2021-Dec 2022

- Lead for Robo-TRACIR project, surgical robot for saving lives of in trauma patients via autonomous needle insertion using robotic ultrasound system (RUS).
- Developed **Kalman Filter-inspired** deep neural network for improving needle segmentation in ultrasound images (ISBI 2024) [7] 
- Designed algorithm for autonomously finding venous regions using **Bayesian Optimization** & force feedback on RUS (ICRA) [1] 

Model Based Reinforcement Learning with Image Observations in Presence of Distractors (PI: [Jeff Schneider](#))

Research Assistant

Pittsburgh, PA

May 2021-Present




- Extended SOTA **model-based reinforcement learning** algorithm **Dream to Control** which learns RL policy using images even in presence of distractors.
- Novel idea of increasing representational power of **variational autoencoder** by fusing autoregressive **video prediction network** and dream to control architecture to handle distractors

ROBOTICS INSTITUTE SUMMER SCHOLAR (RISS) PROGRAM, CMU (PI: Professor [Katia Sycara](#))

Undergraduate Student Intern

Pittsburgh, PA

May 2019-Aug 2019




- Selected amongst 40 students worldwide [2] 
- Developed **heterogeneous multi-agent** task allocation algorithm via **mixed integer optimization** with collision avoidance 
- Scaled multi-agent deep reinforcement learning algorithm for predator prey formation control to more agents via transfer learning 

IIIT DELHI (PI: Professor [Sayan Basu Roy](#) and Professor [P. B. Sujit](#))

Student Researcher

New Delhi, India

May 2018-Dec 2020

- Proposed swarm robotics algorithm for parallel task completion using leader/predator agents (SMC 2019) [3] 
- Designed **first-ever** closed-loop reference model for **distributed systems** (CRM-DMRAC) framework, converges faster to desired trajectory than SOTA; **zero-shot parameter learning** designed based on Lyapunov analysis (L-CSS, ACC 2021) [4] 
- Improved CRM-DMRAC to tackle limited bandwidth multi-agent setting via a novel external input estimation using Dynamic Surface Control and **Cooperative Initial Excitation** (TCNS 2022) [5] 

PROJECTS

Novel Parameter Estimation Algorithm for Time-varying Systems (PI: Professor [Sayan Basu Roy](#))

Pittsburgh, PA

Independent Researcher

Jun 2020-May 2022

- Developed a unified algorithm for time-varying system parameter estimation and tracking using adaptive control and Lyapunov analysis outperforming SOTA.
- First unified algorithm to work for both unknown time-varying parameters and unknown constant parameters [6]

Robotic On-Orbit Satellite Servicing (Northrop Grumman)

Pittsburgh, PA

Graduate Researcher

Feb 2021–May 2021

- Designed inverse dynamics and force based **non-linear controller** for 7-DOF robotic arm to enable minimal disturbance docking

Trajectory Optimization for Spacecraft Rendezvous (PI: Professor [Zac Manchester](#))

Pittsburgh, PA

Graduate Researcher (Course Project)

Feb 2021-May 2021

- Designed an MPC based trajectory **optimization** algorithm with **sequential quadratic programming** for docking in SE(3) space

SKILLS

- Languages & Tools: Python, Pytorch, C++, MATLAB, Simulink, ROS, Gazebo, Mujoco, Julia (intermediate)

PUBLICATIONS

1. [Raghavv Goel](#)*, [Abhimanyu](#)*, [Kirtan Patel](#), [John Galeotti](#), [Howie Choset](#), "Autonomous Ultrasound Scanning with Hybrid Force Controller" *International Conference on Robotics and Automation*, (ICRA 2022)
2. [Raghavv Goel](#), [Jaskaran Singh Grover](#), [Sumit Yi Sha](#), [Katia Sycara](#) "Dynamic Task Allocation Using Multi-Agent Mobile Robots", *Robotics Institute Summer Scholars Journal* (RISS 2019 Journal)
3. [Raghavv Goel](#), [John Lewis](#), [Michael A. Goodrich](#), [P. B. Sujit](#), "Predator & Leader Based Swarm Steering for Multiple Tasks", *International Conference on System, Man, and Cybernetics* (SMC 2019)
4. [Raghavv Goel](#), [Sayan Basu Roy](#), "Closed-loop Reference Model based Distributed MRAC for Multi-agent Systems", *IEEE Control Systems Letters* (L-CSS 2021), *American Control Conference* (ACC 2021) | **Journal**
5. [Raghavv Goel](#), [Tushar Garg](#), [Sayan Basu Roy](#), "Closed-loop Reference Model based Distributed MRAC using Cooperative Initial Excitation and Distributed Input Estimation", *IEEE Transactions on Control of Network Systems* (TCNS 2022) | **Journal**
6. [Raghavv Goel](#), [Sayan Basu Roy](#), "Adaptive Control for Time-varying Systems using Dual Adaptation", *Transaction of Automatic Control* (TAC 2024) | **Journal**
7. [Raghavv Goel](#), [Cecilia Morales](#), [Manpreet Singh](#), [Artur Dubrawski](#), [John Galeotti](#), and [Howie Choset](#). "Motion-Aware Needle Segmentation in Ultrasound Images." *IEEE International Symposium on Biomedical Imaging* (ISBI 2024).
8. [Raghavv Goel](#), [Mukul Gagrani](#), [Wonseok Jeon](#), [Junyoung Park](#), [Mingu Lee](#), [Christopher Lott](#). "Direct Alignment of Draft Model for Speculative Decoding with Chat-Fine-Tuned LLMs." *International Conference on Learning & Representation* ME-FOMO Workshop (ICLR 2024).
9. [Mukul Gagrani](#)*, [Raghavv Goel](#)*, [Wonseok Jeon](#), [Junyoung Park](#), [Mingu Lee](#), and [Christopher Lott](#). "On Speculative Decoding for Multimodal Large Language Models." *Conference on Vision and Pattern Recognition*, eVLM Workshop (CVPR 2024)
10. [Wonseok Jeon](#), [Mukul Gagrani](#), [Raghavv Goel](#), [Junyoung Park](#), [Mingu Lee](#), and [Christopher Lott](#). "Recursive speculative decoding: Accelerating llm inference via sampling without replacement." *International Conference on Learning & Representation* LLM-Agents Workshop (ICLR 2024)