

Raghavv Goel

raghavsyzy@gmail.com | [Google Scholar](#) | [Website](#) | [LinkedIn](#) | (412)-519-8580

EDUCATION

Carnegie Mellon University (CMU) | Pittsburgh, PA | Master of Science in Robotics (MSR) | GPA: 3.95/4.0 | Dec 2022

Thesis: **Autonomous Methods for Robotics Ultrasound Scanning and Needle Segmentation**

Indraprastha Institute of Information Technology Delhi (IIITD) | New Delhi, India | Bachelor of Technology in ECE | (Dept. Rank 1/70) | May 2020

Thesis: **Multi-agent Swarm Robotics Control using Predator and Leader agents**

SKILLS

Languages: Python, C++, MATLAB, Simulink, Julia

Frameworks: PyTorch, ROS, Gazebo, MuJoCo, TensorFlow

Methodologies: Large Language Models (LLMs), Speculative Decoding, KV Cache Optimization, Model-based Reinforcement Learning, Lyapunov Stability Analysis, Adaptive Control, Convex Optimization, Optimal Control

EXPERIENCE – LLMs

Senior Research Scientist | Qualcomm AI Research | San Diego, CA | Jan 2023 - Present |

- Leading research on efficient inference and memory optimization; **filed 24 patents** in 3.5 years across speculative decoding, KV-cache management, diffusion LLMs, and multimodal generation with publications in ICML, Neurips, ICLR.
- **Speculative Decoding:** Developed multiple inference acceleration methods including the first training-free drafter-free N-gram cache-free multi-token prediction (ESP), future-conditioned drafting (ConFu), tree-based decoding (RSD), vocabulary pruning for speculative decoding (VocabTrim), and the first multimodal speculative decoding framework, achieving up to 2.4x generation speedups
- **KV-Cache Optimization:** Developed KeyDiff and CAOTE, including the first closed-form error minimization framework for token eviction, outperforming state-of-the-art methods while reducing memory footprint
- **Discrete Diffusion LLMs:** Pioneers self-speculative block decoding for diffusion LMs (Spiffy) and conducted one of the first comparative analyses of representational capacity between discrete diffusion and autoregressive LLMs.
- **Alignment and Post-Training:** Developed a novel distillation objective combining policy-gradient optimization with total-variation distance, improving gradient variance and alignment quality by 30% over KL-based approaches
- **Systems Optimization:** 3-4 months stint. Designed integer-optimization techniques for DDR memory scheduling, eliminating key memory bottlenecks in deep learning workloads without commercial optimization solvers

EXPERIENCE – ROBOTICS, CONTROLS and DYNAMICAL SYSTEMS

Carnegie Mellon University -- Graduate Research Assistant, Robotics & Control

- **Robotics & Autonomous Systems:** Developed perception, estimation, and control algorithms for robotic systems, including autonomous ultrasound-guided vein mapping using a hybrid force/position controller, and learning-based state estimation for ultrasound needle segmentation
- **Optimal Control:** Built model predictive control (MPC) and trajectory optimization frameworks for spacecraft rendezvous and tracking on SE(3), incorporating nonlinear dynamics, constraints and sequential quadratic programming (SQP)-based optimization

IIITD / Independent Researcher – Adaptive & Distributive Control

- **Adaptive Control Theory:** Developed the first unified dual composite-adaptive control framework for time-varying parameter estimation with rigorous Lyapunov stability guarantees.
- **Multi-Agent Systems:** Developed the first **closed-loop reference model (CRM)**-based distributed MRAC for networked leader-follower systems, establishing stability guarantees, faster convergence and improved transient performance over open-loop approaches.
- **Distributed Learning & Cooperative Adaptation:** Extended distributed adaptive control with cooperative excitation and initialization mechanisms, providing first parameter convergence guarantees using purely local graph information in networked multi-agent systems.

PUBLICATIONS

(Divided to highlight both industry Relevance and Theoretical Depth)

Efficient LLM

- **ESP: Efficient Training-Free Multi-Token Prediction via Embedding Space Probing;** R Goel, et al [ICML 2026, [Paper](#)]
- **ConFu: Contemplate the Future for Better Speculative Sampling;** Z Qin*, R Goel*, et al [ICML 2026, [Paper](#)]
- **KeyDiff: Key Similarity-Based KV Cache Eviction for Long-Context LLM Inference;** J Park, D Jones, M Morse, R Goel, et al [NEURIPS 2025, [Paper](#)]
- **CAOTE: efficient KV-caching via attention output error based token eviction;** R Goel, et al [MemAgents Workshop ICLR 2026, [Paper](#)]
- **VOCABTRIM: Vocabulary Pruning for Efficient Speculative Decoding in LLMs;** R Goel et al [ES-FOMO Workshop ICML 2025, [Paper](#)]
- **Recursive Speculative Decoding: Accelerating LLM Inference via Sampling Without Replacement;** W Jeon, M Gagrani, R Goel, et al [LLM-Agents Workshop ICLR 2024, [Paper](#)]
- **Direct Alignment: Draft Model for Speculative Decoding with Chat-Fine-tuned LLMs;** R Goel, et al [ME-FOMO Workshop ICLR 2024, [Paper](#)]
- **Multi-modal Speculative Decoding: On Speculative Decoding for Multimodal Large Language Models;** M Gagrani*, R Goel* et al [eVLM Workshop CVPR 2024, ORAL, [Paper](#)]

- **Learnable Kalman-filter:** Motion-aware Needle Segmentation in Ultrasound Images, **R Goel**, et al, [ISBI 2024/CVPR 2024 workshop [Paper](#)]

Efficient Diffusion LLM (dLLM)

- **Spiffy:** Multiplying Diffusion LLM Acceleration via Lossless Speculative Decoding; S Agrawal, R Garrepalli, **R Goel**, et al [ICML 2026 FoGen workshop, [Paper](#)]
- A Comparative analysis of **Layer-wise Representational Capacity** in AR and Diffusion LLMs; **R Goel***, R Garrepalli*, et al [Sci4DL, DeLTA workshop ICLR 2026, [Paper](#)]

Control Theory & Robotics (Demonstrating rigorous mathematical foundation in stability analysis and adaptive systems, **all first authors**)

- **Adaptive Systems:** Adaptive Control for Time-Varying Systems using Dual Adaptation, IEEE Transactions on Automatic Control (TAC) 2025 [[IEEE](#), [Arxiv](#)]
- **Multi-Agent Adaptive Systems:** Closed-loop Reference Model based Distributed MRAC for Multi-agent Systems, IEEE Control Systems Letters (L-CSS) & ACC 2021 [[IEEE](#)]
- **Multi-Agent Adaptive Systems 2:** Closed-loop Reference Model based Distributed MRAC using Cooperative Initial Excitation, IEEE Transactions on Control of Network Systems (TCNS), 2022 [[IEEE](#)]
- **Robotics:** Autonomous Ultrasound Scanning with Hybrid Force Controller, ICRA 2022 [[IEEE](#)]
- **Swarm Robotics:** Leader and Predator Agent based swarm steering, IEEE System, Man, and Cybernetics (SMC), 2019 [[IEEE](#)]