Raghavv Goel

raghavsyz@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 Electronics & Communication | (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

Senior Deep Learning Researcher (Systems) | Qualcomm AI Research | San Diego, CA | Jan 2023 - Present

- **LLM Efficiency Optimization:** Leading research on efficient inference and memory optimization, spanning speculative decoding (SPD), KV-cache management, and multi-modal acceleration
- **KV Cache Innovation:** Developed **KeyDiff** (**Neurips 2025**), a training-free eviction heuristic based on key similarity, reducing cache size by **23%** with minimal performance degradation
- **Meta-Heuristic Cache Eviction**: Designed **CAOTE**, utilizing a closed-form solution to optimize eviction error, boosting/outperforming baseline strategies like H2O and SnapKV.
- **Speculative Decoding**: Pioneered **Multi-Modal SPD** (CVPR 2024 eVLM workshop Oral) showing **2.4x speed-up** boost over autoregressive generation,
  - **Recursive Speculative Decoding** (ICLR 2024 LLM-Agents Workshop) using tree-based sampling without replacement to improve acceptance rates, and
  - **VOCABTRIM** (ICML 2025, ME-FoMo Workshop) first paper to improve SPD efficiency by trimming draft model LM-head in principled manner showing **16% memory-bound speed-up over Eagle-2/3**.
  - **ConFu** contemplating future tokens to better guide the draft model generation showing ~10% improvement over EAGLE3 token acceptance and speed-up.
  - **Self-Speculative Decoding:** Proposed **first training-free drafter-free true multi-token prediction** giving 1.6-1.8x acceptance rate (under submission ICML 2026)
- **Model Alignment**: Proposed a novel distillation loss combining policy gradient and **Total Variation Distance (TVD++)** improving alignment by **30%** over KLD baselines (ICLR 2024 ES-FOMO workshop)
- **Accelerating Diffusion LLM:** Proposed first self-speculative block decoding for diffusion LLMs (**Spiffy**). Additionally proposed **Skip to the Good Part**: Representation analysis and layer-skipping for dLLMs. Showcasing how **training objective** can result in representation redundancy and how AR-M initialization can impact dLLM training.
- **Resource-Constrained Optimization:** Solved DDR memory bottlenecks for deep learning models using a novel integer optimization algorithm, achieving optimal scheduling without licensed solvers.

Carnegie Mellon University | Pittsburgh, PA | Graduate Research Assistant (Biorobotics Lab & Auton Lab | Jan 2021-Dec 2022

- **Medical Robotics**: Led the Robo-TRACIR project, surgical robot for saving lives of in trauma patients. Developed a **Kalman Filter-inspired deep neural network** for needle segmentation in ultrasound images (ISBI 2024, CVPR 2024). Developed **force-feedback controller** for 6 DOF robotic ultrasound scanner using **Bayesian Optimization** for autonomous vein mapping (ICRA 2022).
- **Model-Based RL**: Extended "Dream to Control" algorithm by fusing autoregressive video prediction with **variational autoencoders** (VAEs) to handle visual distractors in RL tasks.
- **Control Systems**: Designed an MPC-based trajectory optimization algorithm using sequential quadratic programming for spacecraft rendezvous in SE(3)

SELECTED PUBLICATIONS

(Divided to highlight both industry Relevance and Theoretical Depth)

Generative AI and Efficient ML

- **KeyDiff**: Key Similarity-Based KV Cache Eviction for Long-Context LLM Inference, Neurips 2025, J Park, D Jones, M Morse, **R Goel**, et al [Arxiv]
- **CAOTE**: efficient KV-caching via attention output error based token eviction, **R Goel**, et al [Arxiv]
- **VOCABTRIM**: Vocabulary Pruning for Efficient Speculative Decoding in LLMs, ICML 2025 (ES-FOMO), **R Goel** et al [Arxiv]
- **Recursive Speculative Decoding**: Accelerating LLM Inference via Sampling Without Replacement, ICLR 2024 (LLM-Agents), W Jeon, M Gagrani, **R Goel**, et al [Arxiv]
- **Direct Alignment**: Draft Model for Speculative Decoding with Chat-Fine-tuned LLMs, ICLR 2024 (ME-FOMO), **R Goel**, et al [Arxiv]
- **Multi-modal Speculative Decoding**: On Speculative Decoding for Multimodal Large Language Models, CVPR 2024 (eVLM, Oral), M Gagrani*, **R Goel*** et al [Arxiv]
- **Learnable Kalman-filter:** Motion-aware Needle Segmentation in Ultrasound Images, ISBI 2024/CVPR 2024 workshop, **R Goel**, et al [Arxiv]

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) 2024 [IEEE, Arxiv]
- **Distributed Systems**: Closed-loop Reference Model based Distributed MRAC for Multi-agent Systems, IEEE Control Systems Letters (L-CSS) & ACC 2021 [IEEE]
- **Network Systems**: 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]

Under submission ICML 2026

- **Spiffy: Multiplying Diffusion LLM Acceleration via Lossless Speculative Decoding**, S Agrawal, R Garrepalli, **R Goel**, et al [Arxiv]
- **ConFu: Contemplate the future for better speculative sampling**, Z Qin*, R Goel*, et al
- **Efficient Training-free Multi-token prediction via Embedding Space Probing**, R Goel et al
- **Skip-to-the-Good Part: Representation Structure & Inference Time Layer-skipping in Diffusion vs Autoregressive LLMs**, R Goel*, R Garrepalli* et al