RISHABH JAIN

(646)-484 0525 rj2790@columbia.edu r-rishabh-j.github.io LinkedIn GitHub

EDUCATION

Columbia University, New York **Aug 2025 - Dec 2026**

Master of Science (MS) Computer Science (Machine Learning) | GPA: 3.92/4

- Spring'26: ML Decision Theory, Robot Manipulation, Quantum Computing
- Fall'25: ML, High Performance ML, NLP, Computational Aspects of Robotics

Indian Institute of Technology Ropar **July 2019 - May 2023**

B.Tech (Honors) Computer Science and Engineering, Concentration in Artificial Intelligence

- Data Science, Artificial Neural Networks, Artificial Intelligence, Advanced Computer Vision

PROJECTS AND RESEARCH

NFR Benchmarking in IBM ITBench for IT Automation Agents **Oct 2025 - ongoing**

PyTorch, ITBench, CrewAI, Langfuse | GitHub: ITBench-NFR *IBM Research*

- Co-developing a non-functional requirements (NFR) evaluation framework extending ITBench - defining a two-level taxonomy for agent-specific requirements (cost efficiency, reliability, observability) and instrumenting SRE, CISO and Mini-SWE agents with Langfuse, and vLLM for granular telemetry
- Conducted comparative evaluations across ReAct and Plan&Execute architectures on 15 SRE incidents and 3 CISO scenarios from ITBench using Gemini-2.5-Pro and Qwen3-14B LLMs

Accelerating Speculative Decoding via On-Policy Knowledge Distillation **Oct 2025 - Dec 2025**

PyTorch, Huggingface TRL | github.com/r-rishabh-j/distillSpec, batched_specdec *Columbia University*

- Implemented a speculative decoding library using PyTorch and HF Transformers with prompt batching, non-uniform acceptance length, kv-caching and pruning, and parallel draft verification
- Fine-tuned low-cost draft models from Qwen3, SmolLM families via white-box, token-level On-Policy Knowledge Distillation with larger target models to mitigate 'exposure bias' to accelerate speculative generation
- Benchmarked token and sequence level acceptance rates over Forward KL, Reverse KL and JSD divergence objectives, achieving a max of 5% increase in token acceptance rate after 1 epoch of distillation on GSM8k and 4% on CNN-DM

Viewpoint-Invariant Robot Manipulation via 3D Geometric Priors **Oct 2025 - Dec 2025**

Python, PyTorch, Mujoco, Gymnasium | github.com/r-rishabh-j/3DEgoACT *Columbia University*

- Modified ACT to take PointNet-encoded 3D point-cloud as input tokens along with egocentric 2D features to mitigate inference-time covariate shift from view-point perturbations in imitation learning policies
- Performed ablations to demonstrate that egocentric cues are crucial alongside allocentric 3D features for contact-rich tasks under a fixed training budget
- Demonstrated zero-shot generalization to perturbed viewpoints, achieving ∼70% success rate in scenarios where standard ACT failed by effectively decoupling global geometric structure from view-dependent appearance

Video Transformer Based Multi-view Body Behaviour Recognition **May 2023 - Oct 2023**

Python, PyTorch, Deep Learning, Computer Vision *Monash University & IIT Ropar*

- Built a multi-view feature-fusion pipeline with a finetuned VideoSwin transformer backbone for multi-label classification | GitHub: MAGIC-TBR
- Published MAGIC-TBR: Multi-view Attention Fusion for Transformer based Bodily Behavior Recognition in Group Settings at ACM MultiMedia, 2023
- Published Multi-view Attention Fusion for Explainable Body Language Behavior Recognition at IEEE TAFFC
- Placed 2nd in the ACM MultiMedia 2023 Bodily Behaviour Recognition Grand Challenge | certificate

Spatio-Temporal Hotspot Detection in Microsoft Azure | BTech Capstone **Aug 2022 - Nov 2023**
Java, Python, PostgreSQL, PostGIS *Microsoft & IIT Ropar*

- Formulated a statistical framework to identify spatio-temporal hotspots in **Microsoft Azure** from network autonomous system data from 10+ Indian cities stored in a spatial PostGIS database
- Contributed to implementation of algorithms and database CRUD, creation of synthetic data, and testing on Microsoft's proprietary dataset
- Publication: Periodic Spatio-Temporal Colored Hotspot Detection in Azure Traffic Data; ACIIDS 2025

WORK EXPERIENCE

Software Engineer, Arista Networks **Jul 2023 - Jun 2025**
C, C++, Python, Docker, Software Defined Networking *Bengaluru, India*

- Implemented low-latency data and control plane components in C++ for the EOS software forwarding engine, enhancing throughput and efficiency for high-volume packet processing and telemetry across 16 repositories
- Contributed to scalable state management and synchronization modules in C++ for ICMP, TCP, IPFIX and GRE protocol stacks to build support for a 9× capacity increase in the EOS concurrent packet flow hash table
- Re-wrote flow table scale test libraries with multiprocessing in Python, achieving a 6× gain in eval throughput
- Created an internal RPM build tool to resolve upstream AlmaLinux dependency graphs with Arista patches during a company-wide shift from Perforce mono-repo to a Git multi-repo setup. Streamlined development workflows for 15+ teams

Edison AI Intern, General Electric Healthcare **May 2022 - Jul 2022**
Python, PyTorch, FastAPI, PostgreSQL, Docker *Bengaluru, India*

- Co-developed an end-to-end spatio-temporal patient tracking pipeline using YOLOv5 and PostgreSQL, allowing for the continuous storage and retrieval of patient location data. Deployed it through containerized endpoints
- Ablated YOLOv5 to develop a lightweight model, and fine-tuned it under resource-constraints (5GB GPU) on 30,000 self-annotated images

TECHNICAL SKILLS

Languages: C, C++, Python, Java, PostgreSQL

Tools: Bash, Linux, Git, Perforce, Docker, GCP, Gemini&OpenAI API, Langfuse, vLLM, MuJoCo

Libraries: FastAPI, OpenCV, CUDA, PyTorch, Gymnasium, Huggingface Transformers

OTHER PROJECTS

RFDN Variants: Efficient Image Super-Resolution NTIRE CVPR Challenge **Feb 2023-May 2023**
Supervised by Dr. Abhinav Dhall | *document* *LASII Lab, IIT Ropar*

- Developed efficient **image super-resolution model variants** of the CNN based RFDN baseline
- Studied trade-offs between accuracy and runtime among variants and achieved a superior PSNR on the DIV2K dataset along with a reduced model inference time

Client Selection in Deep Federated Recommender Systems **Oct 2022-March 2023**
Supervised by Dr. Shweta Jain | *document* *Game Theory Lab, IIT Ropar*

- Developed client subset selection strategies to optimize training costs in **federated recommender systems**
- Evaluated the strategies over collaborative filtering based deep recommender systems on MovieLens datasets

Dynamic Planning in Dyna-Q for Faster Training **Sept 2022 - Nov 2022**
Supervised by Dr. Shashi Shekhar Jha | *document* *IIT Ropar*

- Studied impact of a dynamic planning schedule in **DYNA-Q** & **Deep DYNA-Q** reinforcement learning algorithms.
- Evaluated trade-offs in performance and training costs of various schedules on OpenAI Gym environments.