RISHABH JAIN

(646) 484-0525 rj2790@columbia.edu r-rishabh-j.github.io linkedin.com/in/rishabhj11 github.com/r-rishabh-j

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

- Algorithms, Databases, Operating Systems, Data Science, Neural Networks, Advanced Computer Vision

PROJECTS AND RESEARCH

Dexterous Robot Manipulation with 'SpikeATac' tactile feedback **Feb 2026 - ongoing**

PyTorch, ROS, IsaacSim *ROAM Lab, Columbia University*

- Working on dexterous manipulation policies for 4-fingered manipulator with tactile and visual inputs
- Developing real-to-sim pipelines to RL finetune IL policies on tactile feedback for bolt threading

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

Python, PyTorch, Mujoco, Gymnasium | *GitHub: r-rishabh-j/3DEgoACT* *Columbia University*

- Enhanced ACT to fuse 3D point-cloud with egocentric 2D views to mitigate inference-time view-point perturbations in low-cost robot imitation learning based policies
- Performed ablations to demonstrate importance of egocentric-allocentric feature fusion, and demonstrated zero-shot generalization (∼70% success) to perturbed viewpoints in scenarios where baseline failed

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

PyTorch, Huggingface TRL | *GitHub: r-rishabh-j/distillSpec, r-rishabh-j/batched_specdec* *Columbia University*

- Implemented a speculative decoding engine with prompt batching, non-uniform acceptance length, batched verification, kv-cache pruning
- Distilled Qwen3-0.6B and SmolLM-360M drafters from Qwen3-4B and SmolLM-1.7B respecitvely via sequence level white-box On-Policy Knowledge Distillation to align models for accelerating speculation
- Benchmarked token and sequence level acceptance rates over Forward KL, Reverse KL and JS divergence objectives, achieving 5% increase in token acceptance rate after 1 epoch on GSM8k and 4% on CNNDM

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

PyTorch, ITBench, CrewAI, Langfuse, LLMs | *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
- Compared ReAct and Plan&Execute agents on ITBench scenarios using Gemini-2.5-Pro and Qwen3-14B

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 model with a finetuned VideoSwin transformer for a multi-label task
- 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 **Aug 2022 - Nov 2023**
Java, Python, PostgreSQL, PostGIS | *document* *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, synthetic data curation, and eval on Microsoft's proprietary dataset

RFDN Variants: Efficient Image Super-Resolution | **NTIRE CVPR Challenge** **Feb 2023-May 2023**
Python, PyTorch, Computer Vision | *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

COMPETITIONS

- **Feb 2026**: **Won Qualcomm Snapdragon Multiverse Hackathon** by building a multi-device multi-modal AI coding assistant on Snapdragon devices supporting user input beyond text (voice, stylus annotations)
- **Jul 2023**: Runner-up in **ACM MM Grand Challenge** for building a multi-view video classification model

TECHNICAL SKILLS

Languages: C, C++, Python, Java, RISC-V, Bash

Tools: Linux, Git, Perforce, Docker, Google Cloud, Gemini & OpenAI API, Langfuse, vLLM, MuJoCo

Libraries: Numpy, Pandas, FastAPI, OpenCV, CUDA, PyTorch, Gymnasium, Huggingface

WORK EXPERIENCE

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

- Worked on low-level BESS & DPDK C++ modules in EOS software forwarding engine for scaling packet processing throughput and memory access efficiency
- Led creation of stateful bi-directional flow modules in CloudVision IPFIX across ICMP, TCP, GRE stacks
- Contributed in building support for a 9× capacity increase in the EOS concurrent flow hash table. Re-wrote flow table scale test suite with multiprocessing in Python, achieving a 6× gain in evaluation throughput
- Designed configuration CLIs and SysDB agents to support for MSS firewall in EOS network switches
- Created internal RPM build tools to resolve upstream AlmaLinux dependencies with Arista patches during a company-wide shift from Perforce mono-repo to Git multi-repo. Streamlined development workflows for 15+ teams with rapid adoption and contributions within 1 week of release

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

- Developed a computer vision based system for real-time, face-indexed spatio-temporal tracking of admitted patients in hospitals. Deployed it through containerized endpoints in the Edison Health platform
- Finetuned an ablated YOLOv5 model on a dataset of self-curated 30,000 images on a GPU with just 5GB VRAM, taking over 280 GPU hours

PUBLICATIONS

1. Multi-view Attention Fusion for Explainable Body Language Behavior Recognition. *IEEE TAFFC*, 2025
2. MAGIC-TBR: Transformer-based Bodily Behavior Recognition in Group Settings. *ACM Multimedia*, 2023