

# Muhammad Faheem

+92 331-9090179 | ✉ faheemgurkani@gmail.com

in muhammad-faheem | faheemgurkani | faheemgurkani.github.io

## EXPERIENCE

### National Centre for Physics (NCP)

#### Hosted Researcher

Islamabad, Pakistan

July 2024 – July 2025

- Conducted a comparative analysis of classical and deep learning-based object tracking methodologies, benchmarking single-object tracking (SOT) architectures across standardized evaluation protocols in a research-driven environment; designed experiments end-to-end from hypothesis to evaluation
- Researched deep learning-based fault detection and anomaly detection on sensor time-series data for HVAC systems, leading to an accepted IEEE conference paper titled *“Deep Learning-Based Fault Detection Framework for Predictive Maintenance in HVAC Systems”* at the 13th IEEE Conference on Technologies for Sustainability (SusTech 2026), demonstrating experience in dataset curation, reward signal design, model validation, and scientific reporting

### National Aerospace Science & Technology Park (NASTP)

#### AI Engineering Intern

Rawalpindi, Pakistan

June 2025 – August 2025

- Designed and implemented an LLM-assisted OCR pipeline for structured entity extraction, integrating inference serving via a gRPC-based API and deploying the system into an internal contact management workflow
- Built an offline, gRPC-based multi-person behavior analysis system integrating object detection, multi-object tracking (MOT), human pose estimation, and spatio-temporal action recognition; architected the distributed inference pipeline for controlled evaluation and benchmarking across multi-modal inputs

### R7SimSens

#### Research Engineer (Remote)

Virginia, United States

March 2026 – Present

- Designing and maintaining LLM-based pipeline components for autonomous vehicle simulation infrastructure; managing structured data generation workflows at scale across multi-modal inputs in distributed compute environments
- Contributing to multi-agent AI architectures integrating state-of-the-art world foundation models, with focus on robust pipeline design, distributed training data workflows, validation, and deployment across GPU-accelerated compute environments using Docker and cloud infrastructure

## PUBLICATIONS & RESEARCH

### Accepted

- M. Faheem et al., *“Deep Learning-Based Fault Detection Framework for Predictive Maintenance in HVAC Systems,”* 13th IEEE Conference on Technologies for Sustainability (SusTech 2026).

### In Preparation / Active Submission

- M. Faheem et al., *“Per-Layer Clipping and Gradient Heterogeneity in Differentially Private Federated Learning for Speech Foundation Models,”* targeting INTERSPEECH 2026 / ICASSP 2027. Extends Apple NeurIPS 2025 per-layer clipping to WavLM/HuBERT for Speech Emotion Recognition and Speaker Verification; gradient heterogeneity analysis across tasks.
- M. Faheem and M. Qaiser, *“Learning Kill-Chain Phase from Behavioral Context: A Comparison of Sequence Models Under API-Identity Ablation,”* IEEE conference paper. BiLSTM-CRF, Transformer, and hybrid sequence labeling on Avast-CTU CAPE Dataset; deliberate API-identity ablation to simulate zero-day and obfuscated traces.
- M. Faheem et al., *“Meta-Learned SOZ-Aware Adaptive Brain Graph Transformer for Cross-Patient Seizure Onset Prediction,”* targeting NeurIPS / ICLR / AAAI. Adaptive graph learning, meta-learning for unseen-patient generalization, and SOZ/electrode-coordinate clinical metadata integration on the HUP iEEG dataset.
- M. Faheem et al., *“KV Bench: A Unified Evaluation Framework for Online and Offline Evaluation of KV Cache Compression Algorithms in Open-Source Large Language Models.”* Unified interception-and-transformation engine benchmarking TurboQuant, QJL, and RocketKV on Qwen3-1.7B; dual offline/online metrics across 27 GPU sweep jobs on WikiText-2.

## PROJECTS

---

### KV-Cache Interception & Transformation Engine for LLM Inference

- Designed and implemented a modular KV-cache interception and transformation engine for large language models using **PyTorch**, **Hugging Face Transformers**, and **Qwen3-1.7B**, enabling plug-and-play integration of KV-cache compression algorithms without modifying model or evaluation code; developed the **TurboQuant** compression plug-in via Walsh-Hadamard Transform, Lloyd-Max quantization, and QJL residual encoding
- Built a unified evaluation pipeline benchmarking TurboQuant, KIVI, QJL, and RocketKV across perplexity, inference throughput, memory reduction, and compression-fidelity metrics, establishing a reproducible common yardstick for KV compression methods prior to deployment

### Driver Assistance System Using RT-DETR

- Designed a real-time ADAS inference pipeline using fine-tuned RT-DETR, achieving **+18.6% mAP@50** over baseline DETR on dashcam-style data, with end-to-end inference latency of **38–45 ms/frame** on RTX-class GPUs; optimized throughput via batched inference and ONNX export
- Deployed a production-grade inference stack using FastAPI (ASGI) and Next.js, supporting **1080p video streams at 25 FPS** with REST-based analytics endpoints and structured logging for data quality monitoring

### Action Segmentation Using MS-TCN

- Engineered a multi-stage temporal convolutional network (MS-TCN) in PyTorch for frame-level action segmentation, achieving **+13.8% F1@50** over single-stage TCN baselines on long untrimmed video sequences
- Optimized training with mixed precision (AMP) and GPU-aware batching, enabling **sub-4.8 ms/frame inference** and stable convergence on sequences exceeding **10k frames/video**

### BastionFed: Federated Learning & SOC Platform (Healthcare / IoMT)

- Built a full-stack SOC platform (**Next.js 14 + FastAPI/SSE**) with alert triage, incident forensics, federated model zoo, hash-chained audits, and an LLM copilot spanning **5+ analyst workflows**; designed distributed training data pipelines across heterogeneous client nodes
- Trained a distributed federated FV-DNN with multimodal drift detection, achieving **ROC-AUC  $\geq 0.93$**  on global splits; authored system spec over **Firestore/GCP** (SQL, object storage, Redis) and managed infrastructure-as-code for reproducible deployment

### AERIS: Air Emissions Regional Intelligence System (NASA Space Apps Challenge)

- Built a full-stack environmental intelligence platform ingesting **NASA TEMPO** satellite data via the Harmony API, processing multi-gas measurements ( $\text{NO}_2$ ,  $\text{O}_3$ ,  $\text{CH}_2\text{O}$ , PM, Aerosol Index) using **NetCDF4** and **xarray**, with automated spatial clustering (SciPy) for **5-tier** pollution hotspot classification
- Deployed a FastAPI + Next.js inference stack integrating EPA AirNow ground sensors, real-time weather forecasts, and an LLM interpretation service; implemented distributed data movement from satellite API to serving layer, delivering air quality heatmaps and route safety scoring via Cartopy and Leaflet.js

## EDUCATION

---

### National University of Computer and Emerging Sciences

Islamabad, Pakistan

#### B.S. Artificial Intelligence

Aug 2022 – Present

- Honors: Dean's List (Spring 2023, Fall 2025) | GPA: 3.21/4.00
- Relevant Coursework: Reinforcement Learning, Generative AI, Agentic AI, Computer Vision, ANN, Machine Learning, NLP, Data Structures, Algorithms, Digital Image Processing

## SKILLS

---

**Programming & Frameworks:** Python, C/C++, JavaScript, TypeScript, Rust (learning), FastAPI, React/Next.js

**AI & ML:** PyTorch, TensorFlow, Keras, scikit-learn, YOLO, ONNX, TensorRT, Hugging Face, LangChain, LangGraph, vLLM, n8n

**RL & Training:** Policy gradient methods, GRPO, reward modeling, on-policy data collection, distributed training, mixed-precision (AMP), LoRA/QLoRA

**Scientific Computing:** xarray, NetCDF4, SciPy, NumPy, Pandas, Cartopy, Matplotlib, JAX

**Infrastructure & Systems:** Linux/UNIX, Docker, Kubernetes, Git, AWS, GCP, Terraform (IaC), gRPC, Ray, MCP (Model Context Protocol)