Ahmad Mahmood

+41 762283276 | +1 4153362698 | amahmood@student.ethz.ch | github.com/ahmad-573

Education

ETH Zurich Sep 2023 - Jan 2026
MSc Computer Science | GPA: 5.4 / 6 *Zurich, Switzerland*

Lahore University of Management Sciences (LUMS) Sep 2019 - May 2023
BSc Computer Science | GPA: 3.8 / 4 *Lahore, Pakistan*

Experience

Cast Insights Jul 2025 - Present
Machine Learning Engineer *San Francisco, USA*

- Built an **agentic pipeline** using **Pydantic AI** with tools including a PCA-subspace RAG pipeline projecting query embeddings onto the corpus manifold, achieving **NDCG@1 +0.44%, Recall@1 +0.39%** improvement on Natural Questions.
- Built and shipped a **story detection** and **recommendation system** using NER and POS tagging, improving story quality score by **46%**.
- Deployed HNSW index on text embeddings for **91% similarity search speedup**.
- Implemented and maintained an **MCP server** using the **FastMCP** SDK and OAuth 2.0 authentication flow for client integrations.

Disney Jan 2025 - Sep 2025
Machine Learning Engineer *Zurich, Switzerland*

- Implemented a diffusion-SSM hybrid model for image deblurring by integrating diffusion generative priors with Mamba state-space sequence models.
- Built a custom hidden-state initialization for Mamba selective scan kernels that injects diffusion priors at **2x overhead vs. 10x** for the naive approach, without modifying CUDA kernels.

INSAIT Mar 2024 - Oct 2024
Computer Vision Engineer (Research Intern) *Sofia, Bulgaria*

- Built InTraGen, a trajectory-conditioned diffusion transformer (DiT) video generation pipeline achieving **MTEM 9.40% vs. 22.17%** baseline and **FID 26.37 vs. 40.65**; outperformed LUMA, I2VGen-XL, SEINE. Accepted to **ICCV 2025**.
- Designed and implemented MTEM, a trajectory evaluation metric using bipartite graph matching (Hungarian algorithm); adopted as the primary benchmark for the paper.
- Engineered the ViN dataset pipeline - **50K synthetic videos** across 4 physics-simulated interaction subsets using Blender and Unity.

Mohamed bin Zayed University of AI Mar 2022 - Dec 2022
Research Engineer (Intern) *Abu Dhabi, UAE*

- Implemented temporal prompt optimization on frozen vision models (ViTs, DINO, CLIP) to boost cross-modal adversarial transferability (**ICLR 2023**); built VURF, an LLM-based video understanding pipeline for QA and reasoning (**NeurIPS 2024**).

Projects

3D Scene Reconstruction from Multi-View Images | *PyTorch, Open3D* 2025

- Built a 3D reconstruction pipeline from multi-view images using differentiable rendering, camera pose estimation, feature matching, and volumetric rendering for novel view synthesis.

Conditional Diffusion Models for Image Inpainting | *PyTorch, DDPM* Jun 2025

- Implemented a class-conditional diffusion model (multi-scale UNet) from scratch; applied to inpainting by conditioning the reverse diffusion process on partially observed pixels.

Neural Network Robustness Verification (DeepPoly) 2024

- Implemented a neural network verifier using DeepPoly convex relaxation to certify adversarial robustness on MNIST and CIFAR-10; analyzed tightness vs. computational cost trade-offs.

Selected Publications

Boosting Adversarial Transferability using Dynamic Cues - Mahmood et al., **ICLR 2023**
InTraGen: Trajectory-controlled Video Generation for Object Interactions - Mahmood et al.
VURF: General-purpose Reasoning and Self-refinement Framework for Video Understanding - Mahmood et al., **NeurIPS 2024 Workshop**

Honors and Awards

International Mathematics Olympiad (IMO) 2019 Top **6** in Pakistan (Bath, UK) **Dean's Honour List** Years 1-3 at LUMS (cGPA **3.85-3.93**/4.0)
Google Kickstart 1st place (Pakistan), multiple rounds

Technical Skills

Languages: Python, C/C++, JavaScript, Haskell, LaTeX
Frameworks: PyTorch, Diffusion Models (DDPM/DiT/Score-based), LangChain, LangGraph, Flask, ReactJS, FastMCP
Infrastructure: AWS, Google Cloud, CUDA, Git, Docker