# Kazi Hasan Ibn Arif

403 Progress St NE, Blacksburg, VA-24060, United States
5404496919 hasanarif@vt.edu LinkedIn Scholar

PhD student in Machine Learning & Systems with a peer-reviewed publication record at top ML conferences. Research focuses on optimizing the training and inference of multimodal LLMs, specializing in techniques like efficient self-attention, token dropping, layer skipping, sparsity, and quantization. Prior experience as an ML Engineer at a Fortune 500 company developing and deploying business recommendation systems.

## EDUCATION

**Virginia Tech** **Aug 2023 – Present**
*PhD in Computer Science (CGPA: 3.95/4.00)* *Blacksburg, Virginia, USA*

**Bangladesh University of Engineering and Technology** **Feb 2017 – May 2022**
*Bachelor's in Computer Science and Engineering* *Dhaka, Bangladesh*

## WORK EXPERIENCE

**Graduate Research Assistant** **Aug 2023 – Present**
*Virginia Tech* *Blacksburg, VA, USA*

- We are enhancing the inference efficiency of Multimodal LLMs for edge GPU. This project is funded by NSF (Awd: 2315851) and supervised by Dr. Bo Ji.

**Machine Learning Engineer** **May 2022 – Aug 2023**
*IQVIA* *USA (Remote)*

- I performed research and development for the Next Best - Recommendations Platform of IQVIA Orchestrated Analytics. Besides, I gained industry expertise deploying cloud-based ML pipelines using Kubeflow.

## PUBLICATIONS

**HiRED: Attention-Guided Token Dropping for Efficient Inference of High-Resolution Vision-Language Models**
*Accepted at AAAI 2025 Main Track. (Preprint)*

- We developed a method to improve inference efficiency of multimodal (image + text) large language models (LLMs) by strategically dropping visual tokens during image encoding, addressing the high computational cost of transformers.
- We speed-up the inference by 4.7× and reduce latency by 78% and cut GPU memory usage by 14% on LLaVA-Next-7B with minimal accuracy impact: -1% on VQAv2, -3% on TextVQA, and 0% on ScienceQA on a Tesla P40 GPU.

**Equitable Skin Disease Prediction Using Transfer Learning and Domain Adaptation**
*Accepted at AAAI Symposia 2024. (Manuscript)*

- We addressed the color biases in existing dermatology models, which often underperform on darker skin tones, by leveraging transfer learning with diverse image domains and enhancing performance through domain adaptation with datasets like HAM10000.
- We have achieved state-of-the-art performance on the Diverse Dermatology Images (DDI) dataset on both underrepresented and common skin tones.

**Hybrid Deep Learning Model for Bangla Parts of Speech Tagging**
*Accepted at INCET 2021 (Manuscript)*

- We developed a hybrid model for Bangla parts-of-speech tagging, combining CNN and BiLSTM parallel input networks with a CNN output network to capture both local and global textual features.
- We address the challenges of limited data availability of an underrepresented language like Bangla and significantly outperformed previous studies in the field. I presented the paper at the conference.

## TECHNICAL SKILLS

**Languages:** Python, C, C++, Java, Shell

**Machine Learning and Frameworks:** PyTorch, Huggingface-transformers, vLLM

**Systems and Cloud:** Linux, CUDA, Git (GitHub, GitLab), Docker, Kubeflow, AWS S3

**Databases:** Oracle, PostgreSQL, MongoDB

## LEADERSHIP

**Secretary, CS@VT Graduate Council (2024-2025)**

- I am an elected Secretary to represent the interests of around 400 graduate students in the Dept. of CS@VT.
- I organize faculty meetings, social events, and job fairs to enhance engagement and growth opportunities.
- I arrange communication between students and faculty to address feedback and foster transparency.

## AWARDS AND SCHOLARSHIPS

**Cyber Innovation Scholars Program 2024:** Awarded a $2000 grant for selection into the CCI SWVA Cyber Innovation Scholars Program.

**Fusemachines AI Fellowship 2022:** I was selected for the year-long fellowship sponsored by H&M, and received best presentation award in the Machine Learning course.

**Dean's List Award (Senior Year):** Received for achieving honors grades in consecutive semesters.

**Admission Test Scholarship:** Awarded for securing $72^{nd}$ place (top 1%) in the 2016 undergraduate admission test at the top engineering school in Bangladesh.

**Bangladesh Physics Olympiad:** I ranked in the top 20 in the divisional round and qualified for the national level.

## PROJECTS

A comprehensive list is available here: GitHub Link

**HiRED-LLaVA-Next** | Link | PyTorch, Huggingface Transformer, Python **2024**

- We speed-up the inference of LLaVA-Next by 4.7x, reduce response latency by 78%, and cut the GPU memory usage by 14% on an NVIDIA TESLA P40 without sacrificing much of its multimodal tasks accuracy.

**Fix In-Context Hallunication of LLaVA** | Link | Python, PyTorch, Huggingface Transformer **2024**

- We mitigate in-context hallucination by 46% (CHAIR score) of Multimodal-LLM like LLaVA by intervening its self-attention and adjust the attentions of visual and text tokens in the LLM generation phase.

**A C++ implementation of Rasterization and Ray Tracing Algorithm** | Link | OpenGL, C++ **2021**

- Implemented Phong illumination, ray-object intersection, multi-level reflections, and texture mapping to render realistic scenes from scratch.

**AI-Enabled Lines of Action Game** | Link | Java, JavaFX **2020**

- Developed a heuristic for an AI-enabled board game with a GUI built using JavaFX.

**CPP Compiler** | Link | Yacc, Lex, C **2019**

- Developed a subset of a C compiler with Lexical, Syntax, and Semantic Analysis, including Intermediate Code Generation. Generated DAGs and TAC and converted them to 8086 assembly code.