

HASAN ARIF

403 Progress St NE, Blacksburg, VA-24060, United States

☎ 5404496919 ✉ hasanarif@vt.edu 🔗 LinkedIn 🎓 Scholar 🌐 Webpage

EDUCATION

Virginia Tech, Blacksburg, Virginia, USA *Aug 2023 – Present*
PhD Student in Computer Science Advised by Dr. Bo Ji.
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh *Feb 2017 – May 2022*
Bachelor's in Computer Science and Engineering

WORK EXPERIENCE

Kaliber Labs, San Francisco, California, AI Researcher Intern *May 2025 – Aug 2025*
RAG-assisted conversational pipeline for home-care robots and porting various efficient GenAI models on base station servers.
SNAIL Lab (Virginia Tech), Blacksburg, Virginia, Graduate Research Assistant *Aug 2023 – Present*
System/Algorithmic Optimization of LLM/LMM Inference
IQVIA, USA (Remote), Machine Learning Engineer *May 2022 – Aug 2023*
Research and Development of AI-driven recommendation engine

PUBLICATIONS

[**AAAI 2025**] **Kazi Hasan Ibn Arif**, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, Bo Ji, “HiRED: Attention-Guided Token Dropping for Efficient Inference of High-Resolution Vision-Language Models”, *Proceedings of the AAAI Conference on Artificial Intelligence* [Paper] [Code]
[**CVPR 2024 Workshop**] **Kazi Hasan Ibn Arif**, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, Chris Thomas, “Fixing Imbalanced Attention to Mitigate In-Context Hallucination of Large Vision-Language Model”, *Proceedings of 2025 CVPR workshops* [Paper] [Code]
[**AAAI 2024 Symposia**] Sajib Acharjee Dip, **Kazi Hasan Ibn Arif**, Uddip Acharjee Shuvo, Ishtiaque Ahmed Khan, Na Meng, “Equitable Skin Disease Prediction Using Transfer Learning and Domain Adaptation”, *Proceedings of the AAAI Symposium Series* [Paper] [Code]
[**INCET 2021**] Muntasir Hoq, **Kazi Hasan Ibn Arif**, Mohammed Nazim Uddin, “Local and Global Feature Based Hybrid Deep Learning Model for Bangla Parts of Speech Tagging.”, *2021 2nd International Conference for Emerging Technology (INCET)* [Paper]

TECHNICAL SKILLS

Languages: Python, C, C++, Java, Shell
Machine Learning and Frameworks: PyTorch, Huggingface-transformers, Triton, vLLM, llama.cpp
Systems and Cloud: Linux, CUDA, Git (GitHub, GitLab), Docker, Kubeflow
Databases: Oracle, PostgreSQL, MongoDB

LEADERSHIP AND SERVICES

Secretary, Computer Science Graduate Council 2024-2025 at Virginia Tech

I am elected as Secretary to represent 400+ graduate students and manage active communication between students and authority within department and beyond

Reviewer, ICLR 2025, CVPR 2025 Workshops

Workshop on Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI

Student Scholar and Volunteer, AAAI 2025, Philadelphia, Pennsylvania, USA

AWARDS AND SCHOLARSHIPS

CCI Cyber Innovation Scholar: Selected as CCI SWVA Cyber Innovation Scholar and awarded \$2000 grant

Best Presentation Award: Received best project presentation in the Machine Learning program offered by Fusemachines Inc in partnership with H&M Group.

Fusemachines AI Fellowship 2022: 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 72nd place (top 1%) in the 2016 undergraduate admission test at the top engineering school in Bangladesh

Bangladesh Physics Olympiad: Ranked 17th in the divisional round and qualified for the national level

PROJECTS

Full list is available here:  GitHub Link

HiRED-LLaVA-Next, [Link](#) | PyTorch, Huggingface Transformer, Python

Speeding-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 LVLM Hallucination, [Link](#) | Python, PyTorch, Huggingface Transformer

Mitigating 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.

Rasterization and Ray Tracing in C++, [Link](#) | OpenGL, C++

Implementing Phong illumination, ray-object intersection, multi-level reflections, and texture mapping to render realistic scenes in C++ without using any library

Lines of Action Game with AI, [Link](#) | [Demo](#) | Java, JavaFX

AI-powered Lines of Action board game using JavaFX, implementing Minimax with Alpha-Beta pruning and heuristic-based move evaluation.

CPP Compiler | [Link](#) | Yacc, Lex, C

A fully functional C++ compiler with Lexical, Syntax, and Semantic Analysis, including Intermediate Code Generation. It generates DAGs and TAC from C++ and converts into x86 assembly code.