Md Khalequzzaman Sarker Likhon

Dhaka, Bangladesh · khalequzzamanlikhon@gmail.com · +880 1711 906648
khalequzzamanlikhon.github.io · github.com/khalequzzamanlikhon · linkedin.com/in/khalequzzaman-likhon ·
ORCID: 0009-0002-7048-0727

Research Interests

Multimodal learning and video understanding, with an emphasis on **reliability**: how vision-language and spatiotemporal models behave under real-world distribution shift, how to calibrate them and make them abstain when unsure, and how to keep multimodal and LLM systems grounded in verifiable evidence. Application areas include human action and anomaly recognition, safety-critical perception, and medical imaging.

Education

Ahsanullah University of Science and Technology Apr 2016 – Jan 2021
B.Sc. in Computer Science and Engineering Dhaka, Bangladesh

- CGPA: 2.74 / 4.00
- **Thesis:** *Design of an Arrhythmia Classification Algorithm Using 2-D Convolutional Neural Networks*
- **Relevant coursework:** Artificial Intelligence, Pattern Recognition, Digital Image Processing, Computer Graphics, Soft Computing, Data Structures, Algorithms

Research Experience

Applied ML Research, Accelx Inc. Jan 2025 – Present
Machine Learning Engineer (safety-critical perception and language systems) Dhaka, Bangladesh

Zero-shot vision-language recognition under deployment shift

How far can a frozen CLIP model go on rare safety events (falls, fights) with no labeled video, and how does it fail on real CCTV?

- Built fall and violence recognizers on OpenCLIP ViT-L/14. Each frame is scored against a bank of contrastive text prompts, and a decision rule combines group scores (a max/mean mix) with a threshold and a margin calibrated by sweeping F1 on labeled deployment frames.
- Studied the failure modes in production footage: overhead and fisheye viewpoints, cast shadows, reclining or seated postures, and close physical contact such as hugging. To address them, I mined 80+ hard-negative prompts directly from observed false positives. This turns an error analysis into a training-free correction loop.

Spatiotemporal and pose-based violence recognition

Which works better for detecting fights in surveillance video: appearance-based spatiotemporal features or explicit human-pose dynamics?

- Compared three model families on RWF-2000:
  - a Kinetics-pretrained R(2+1)D-18 network
  - a BERT-style Transformer over per-frame pose-dynamics features: YOLO11-Pose keypoints for up to 4 people, centroid-matched velocities, elbow angles, and an inter-person contact cue
  - a two-stream model in which pose features attend to 3D-CNN RGB features through cross-attention
- Investigated motion-aware clip mining (dense optical flow plus pose heuristics) to reduce label noise from video-level annotations, and evaluated clip-level predictions over full videos.

Grounded and abstaining multimodal and LLM systems

How can small local models (3B-parameter LLMs, 2B VLMs) be made trustworthy enough for high-stakes decision support?

- Designed a pipeline in which the LLM only writes prose from facts computed deterministically in Python. Every finding must cite its source passage, and a word-overlap test checks those citations. I set the test's threshold from observed failures: a misattributed citation shared about 17% of its words with the source, while correct citations shared close to 100%.
- Built a vision-language threat-analysis cascade: OCR, then tiled CLIP symbol detection, then a VLM only when needed. Every extracted finding is checked against a curated knowledge base, so hallucinated content cannot affect the score. Confidence is calibrated, and weak evidence is handed off to human review.
- Evaluated generated text with an LLM judge, ROUGE-L and fact recall. In a separate RAG study on financial filings, a general-domain cross-encoder reranker *lowered* faithfulness from 0.93 to 0.77, which suggests that reranking is sensitive to domain shift.

Manuscripts and Reports

(* equal contribution)

- R. Tanvir*, M. K. S. Likhon*, et al. *Joint Prediction of Ultra-High-Performance Concrete Properties Using a Feature-Tokenizer Transformer with Masked Multi-Task Learning.* Submitted to *Applied AI Letters* (Wiley), 2026. Under review.
  Jointly predicts multiple concrete properties, using masked multi-task learning to train on partially labeled samples.
- M. K. S. Likhon*, M. M. Sarker*. *Out-of-Focus Is Not Obliteration: A Perturbation Analysis of Fingerprint Alteration Detectors on Synthetic Benchmarks.* Manuscript in preparation, 2026.
  Shows that detectors reaching 0.998 macro-F1 on SOCOFing learn editing traces, not alteration.
- M. K. S. Likhon*, M. M. Sarker*. *Does Geometry Help? A Leakage-Audited, Negative-Result Study of Cross-Attention Fusion and Taxonomy-Consistency Loss for Post-Earthquake Damage-Type Classification.* Manuscript in preparation, 2026.
  Shows that geometry fusion performs no better than a shuffled-geometry control on PEER Hub ImageNet.

Selected Independent Projects

- **DocuRoute: agentic retrieval for financial documents.** An LLM router, grounded in the extracted table schemas, sends each question to hybrid sparse–dense retrieval or to sandboxed text-to-SQL. Answers carry citations. I implemented the evaluation metrics (context precision and recall, faithfulness, relevancy) and used them to choose the retrieval configuration.
- **Research→Code: multi-agent program synthesis.** Researcher, coder and reviewer agents (LangGraph, with tools served over MCP). Generated code is tested in a sandbox, retries are bounded by deterministic routing, and a human approves results before anything is written.
- **Sentinel: multi-camera video anomaly detection.** Detection and tracking (YOLOv8 + BoT-SORT) feed persistence-gated rules for falls, loitering, abandoned objects and wrong-way motion. An LSTM trajectory autoencoder flags patterns that no rule covers.
- **SentinelRAG, DocuVision, voice-dub.** A self-correcting RAG pipeline with hallucination auditing; a classical-CV document layout and table parser; and a speaker-cloned, time-aligned dubbing pipeline (Whisper, pyannote, NLLB, XTTS-v2).
- **Earlier vision work.** Pneumonia detection with an attention-weighted DenseNet121 + EfficientNet ensemble (95% accuracy); semantic segmentation on CamVid comparing FCN, attention U-Net and DeepLabV3+; Faster R-CNN on PASCAL VOC 2012; fingerprint forgery detection on SOCOFing.

Technical Skills

Research tools	PyTorch, TensorFlow/Keras, Hugging Face, OpenCLIP, Ultralytics, scikit-learn, OpenCV, W&B
Methods	Detection, pose estimation and tracking; video classification; vision-language models; transformer fine-tuning; RAG and LLM agents; evaluation design
Engineering	Python, C/C++, SQL, Bash; FastAPI, Docker, TorchServe, Redis, Qdrant, Git, Linux

Honors and Awards

Education Board Scholarship, Secondary School Certificate (SSC) 2012
Education Board Scholarship, Junior School Certificate 2009
3rd Place, District Astro Olympiad 2010

Teaching, Service and Activities

Home tutor for secondary and higher-secondary students (grades 10–12) 2017 – 2022
Organizing Committee, AUST Codeware (programming contest) 2019
Competitive programming: 100+ problems (Beecrowd, Codeforces, CodeChef)

Certifications

Machine Learning – Coursera (2023) · Neural Networks and Deep Learning – Coursera (2021)

Languages

Bengali (native) · English (full professional proficiency)