

Mohammad Nur Hossain Khan

Worcester, MA | mkhan@wpi.edu | (774) 519-0452 | [LinkedIn](#) | [Website](#) | [GitHub](#)

SUMMARY

Late-stage PhD candidate (defense expected 2026) in ECE at Worcester Polytechnic Institute, advised by Dr. Bashima Islam. Research spans **audio large language models** and **audio understanding** and **self-supervised, multimodal learning for noisy real-world physiological time-series** (IMU, ECG/HRV, PPG/IBI, accelerometer, and audio biosignals). First-author work at **EMNLP, Interspeech, ACM IMWUT, IEEE BSN, and IEEE/ACM CHASE** on audio-LLMs (RAVEN, QUADS), wearable health methods, and self-supervised models; co-author on Owl (ICLR 2026), a geometry-aware spatial audio LLM. Contributions include multi-speaker ASR-style diarization, speaker-role classification, non-speech audio event detection, **audio moment retrieval** on long-form recordings, **graph-based self-supervised modeling** for irregular physiological signals, a deployed **commercial mindfulness app (EQUA)** validated in a 42-user study, and **TinyML / on-device inference** (UniT, EfficientMic). Comfortable training and deploying large multimodal models on multi-GPU A100 clusters, with an end-to-end track record from method to product to user evidence.

EDUCATION

Ph.D. in Electrical & Computer Engineering

Aug 2022 – Present

Worcester Polytechnic Institute — Advisor: Dr. Bashima Islam — defense expected 2026. Focus: audio-LLMs, self-supervised learning, multimodal and time-series methods for wearable physiological data.

M.S. in Electrical & Computer Engineering

Conferred May 2025

Worcester Polytechnic Institute

SELECTED RESEARCH

Confidence-Guided Retrieval Refinement for Audio Moment Retrieval

2025 – present

Lead author (in submission)

- Built a retrieve–rerank framework for localizing natural-language queries in long-form (5-min) YouTube audio: cross-modal retriever (audio + text encoders, cosine similarity) returns top- K candidate temporal segments; a second-stage reranker selectively refines them.
- Introduced a **confidence-guided reranking gate**: replaces the top-1 retrieval only when the reranker’s confidence exceeds threshold τ , avoiding destructive corrections on locally ambiguous (IoU-overlapping) candidates.
- Added **retrieval-grounded span refinement**: refined boundaries are constrained by an $\text{IoU} \geq \delta$ overlap criterion with the retrieved candidate, preventing hallucinated span generation.
- Optimized the reranker with **Direct Alignment Preference Optimization (DAPO)** over preferred/rejected candidate pairs.
- Evaluated on **CASTELLA** (1,347 test queries; median 4 s gold spans); retriever pre-trained on Clotho-Moment (51,240 synthetic mixtures) and fine-tuned on CASTELLA. Final system improves over published UVCOM by **+2.82 R1@0.5, +1.97 R1@0.7, and +1.75 mAP**.

Multi-task Audio LLM for Long-Form Home Recordings — Speaker Diarization, Vocalization Classification, and Captioning

2025 – present

Lead author (in submission); LB-LLM

- Built an audio large language model that jointly performs **speaker counting**, **fine-grained speaker-role diarization** (infant / parent / sibling / non-family), and **non-speech vocalization labeling** on long-form home audio at **0.1-second resolution** — a setting where standard audio-LLMs trained on web caption data degrade sharply.
- Trained an **audio spectrogram transformer** for non-speech audio classification across an infant-centric soundscape, combining in-house recordings (Littlebeats wearable) with public datasets including AudioSet.
- Constructed a 5–30 s audio captioning corpus and a stage-wise reasoning dataset for caregiver–infant interaction; implemented **GRPO post-training with the VERL framework** to improve audio-grounded reasoning and instruction-following over the base audio-LLM.

Owl: Geometry-Aware Spatial Reasoning for Audio Large Language Models

2025

Co-author, ICLR 2026 (Biswas, Khan, Islam)

- Co-developed **SAGE**, a geometry-aware binaural audio encoder trained with auxiliary room-impulse-response prediction (panoramic depth + simulated RIRs) so the model learns direct-to-reverberation ratio, RT60, and room layout cues. At inference, only binaural audio is needed.
- Co-built **Owl**, an audio LLM integrating SAGE with a Q-Former projector and LLaMA-2-7B backbone, trained with a three-stage curriculum (perception → relational geometry → chain-of-thought).
- Released **BiDepth**, a 1.1M-QA-pair dataset of binaural audio + RIRs + panoramic depth. SAGE reduces mean DoA error by 11° on SpatialSoundQA; Owl improves spatial QA accuracy by up to 25% over BAT (SOTA baseline).

RAVEN: Query-Guided Representation Alignment for Audio + Video + Sensor QA

2024 – 2025

First author, EMNLP 2025

- Proposed **Query-Aligned Representation of Tokens (QuART)**, a query-conditioned cross-modal alignment mechanism that selectively fuses audio, video, and embedded-sensor features for question answering — robust to noisy and missing modalities, directly applicable to multi-stream wearable health monitoring.
- Designed a three-phase mismatch-aware training framework: modality-specific language-aligned pretraining → joint token–query alignment → discrepancy-aware fine-tuning.
- Curated a **300k-pair multimodal QA benchmark** from Epic-Kitchens-100 and Ego4D; integrated audio, video, and embedded-sensor encoders within a multimodal LLM architecture.

QUADS: Quantized Distillation Framework for Efficient Speech Language Understanding

2024 – 2025

First author, Interspeech 2025

- Designed a joint **knowledge-distillation + quantization** pipeline for speech language understanding, compressing a large teacher into a deployable student with minimal accuracy loss.
- Targets the model-quality / inference-speed / memory-footprint trade-off central to shipping audio understanding models in production and to wearable health-audio deployment.

Mindfulness Outcome Estimation from Smartphone IMU — Deployed in EQUA App

2024 – 2025

First author, ACM IMWUT 2025; follow-up submitted to IMWUT 2026

- Built a statistical-signal-processing pipeline that estimates respiration rate from smartphone accelerometer data with **1.6 breaths/min mean error** — robust enough to ship in a commercial mindfulness app.
- Trained a deep model that predicts session-level mindfulness outcome from IMU motion with **F1 80–84%**, interpreting changes in body stillness as a proxy for meditation skill progression.
- Integrated mindfulness score, respiration chart, and respiration statistics into **EQUA**, a commercial mindfulness meditation app shipped to end users.
- Led a **42-user study** to validate system usability and user engagement with biosignal feedback — an end-to-end loop from algorithm → deep model → product integration → user evidence.
- Follow-up work (*MindfulBreath*, submitted IMWUT 2026): interpretable mindfulness outcome prediction from smartphone IMU respiration.

ECG / IBI Foundation Model — Graph-Based Self-Supervised Pretraining for HRV

2025 – present

Lead author, in preparation

- Designed a **graph neural network** that represents inter-beat-interval (IBI) sequences as physiologically-featurized graphs: 10 HRV-derived per-beat features (including pNN50 / RMSSD proxies), MAD-normalized for cross-subject invariance, with **K-NN similarity edges** and temporal edges built dynamically each forward pass.
- Built a hybrid SSL objective combining **masked-node reconstruction** (local rhythmic context) with **BYOL contrastive loss** under realistic augmentations (jitter, beat dropout, scale drift) — explicitly designed for ECG’s irregular beat sequences, where fixed-rate methods (wav2vec2 / HuBERT) fail.
- Targets downstream arrhythmia detection, autonomic state estimation, and stress/recovery — directly relevant to wearable HRV-based health inference.

Infant Behavior & Parent–Infant Interaction Analysis from Wearable Multimodal Sensors

2023 – present

First author — IEEE/ACM CHASE 2024, IEEE BSN 2024; multiple in-progress

- Building a multimodal framework that analyzes parent–infant interaction from synchronized **audio, ECG, and IMU** streams collected via the in-house Littlebeats chest-worn wearable.
- **Sound Tagging in Infant-Centric Home Soundscapes (CHASE 2024, first author)**: trained an audio spectrogram transformer for non-speech audio event detection across infant-centric environments, combining Littlebeats home recordings with AudioSet to enhance generalization.

- **InfantMotion2vec (IEEE BSN 2024, first author):** unsupervised **contrastive-learning** model that learns infant pose and motion embeddings from unlabeled chest-IMU data — a concrete instance of SSL on noisy, asynchronous, real-world wearable time-series.

LLaSA: Multimodal LLM for Human Activity Analysis from Wearable & Smartphone Sensors

2024 – 2025

Co-author, *UbiComp Companion 2025*

- Co-developed a multimodal LLM that ingests IMU and smartphone sensor streams alongside language for activity and motion analysis — a key building block for conversational health assistants.
- Contributed to **SensorCaps** (26,288 IMU-derived activity narrations) and **OpenSQA** (257,562 instruction-following QA pairs) for sensor-grounded language modeling.

TinyML & On-Device Audio / Sensing Inference — UniT, EfficientMic

2024 – 2025

Co-author (*arXiv 2025, BuildSys 2025*)

- **UniT (arXiv 2025):** contributed to scalable unstructured **inference-time pruning** for MAC-efficient neural inference on **microcontrollers (MCUs)**, enabling deployable audio/sensing models under tight compute budgets.
- **EfficientMic (BuildSys 2025):** co-author on adaptive acoustic sensing using a **single microphone** for smart-infrastructure monitoring; closes the loop from low-power signal acquisition to on-device classification.

SELECTED PUBLICATIONS

1. **Khan, M. N. H.**, Biswas, S., & Islam, B. *RAVEN: Query-guided Representation Alignment for Question Answering over Audio, Video, Embedded Sensors, and Natural Language*. EMNLP 2025.
2. **Khan, M. N. H.**, Biswas, S., & Islam, B. *QUADS: Quantized Distillation Framework for Efficient Speech Language Understanding*. Interspeech 2025, pp. 4098–4102.
3. Biswas, S., **Khan, M. N. H.**, & Islam, B. *Owl: Geometry-aware Spatial Reasoning for Audio Large Language Models*. ICLR 2026.
4. **Khan, M. N. H.**, et al. *Mindfulness Meditation and Respiration: Accelerometer-based Respiration Rate and Mindfulness Progress Estimation to Enhance App Engagement and Mindfulness Skills*. ACM IMWUT, vol. 9, no. 3, pp. 1–32, 2025.
5. **Khan, M. N. H.**, et al. *MindfulBreath: Interpretable Mindfulness Outcome Prediction from Smartphone IMU Respiration*. Submitted, ACM IMWUT 2026.
6. Biswas, S., **Khan, M. N. H.**, Colwell, A., Adiletta, J., & Islam, B. *LOCUS: Localization with Channel Uncertainty and Sporadic Energy*. 22nd Int'l Conf. on Embedded Wireless Systems and Networks (EWSN), 2025.
7. Imran, S. A., **Khan, M. N. H.**, Biswas, S., & Islam, B. *LLaSA: A Multimodal LLM for Human Activity Analysis through Wearable and Smartphone Sensors*. UbiComp Companion 2025.
8. **Khan, M. N. H.**, Li, J., McElwain, N. L., Hasegawa-Johnson, M., & Islam, B. *Sound Tagging in Infant-Centric Home Soundscapes*. IEEE/ACM CHASE 2024, pp. 142–146.
9. **Khan, M. N. H.**, McElwain, N. L., Hasegawa-Johnson, M., & Islam, B. *InfantMotion2vec: Unlabeled Data-Driven Infant Pose Estimation Using a Single Chest IMU*. IEEE BSN 2024, pp. 1–4.
10. Adiletta, J. T., Reynolds, M. C., **Khan, M. N. H.**, Fang, S., & Islam, B. *EfficientMic: Adaptive Acoustic Sensing with a Single Microphone for Smart Infrastructure*. BuildSys 2025, pp. 262–265.
11. Neth, A., **Khan, M. N. H.**, Biswas, S., Salekin, A., Islam, B., et al. *UniT: Scalable Unstructured Inference-Time Pruning for MAC-Efficient Neural Inference on MCUs*. arXiv:2507.07885, 2025.
12. **Khan, M. N. H.**, & Fattah, S. A. *Surface EMG Based Basic Hand Movement Detection Using Singular Value Decomposition*. IEEE BECITHCON 2019, pp. 97–102.
13. **Khan, M. N. H.**, Biswas, S., & Islam, B. *Confidence-Guided Retrieval Refinement for Audio Moment Retrieval*. In submission, 2026.

SKILLS

Audio & speech: multi-speaker ASR / diarization, speaker-role classification, non-speech audio event detection, **audio moment retrieval** (retrieve–rerank, DAPO preference optimization, IoU-grounded span refinement), audio-LLMs (LLaMA-2 + Q-Former, audio spectrogram transformer), binaural / spatial audio, Mel-spectrogram + IPD features, AudioSet pipelines.

Physiological signals: ECG / HRV / IBI, PPG, accelerometer (IMU), respiration (accelerometer-derived), surface EMG, audio biosignals, infant chest-worn wearable streams (Littlebeats), multi-sensor synchronization and cleaning.

Self-supervised & graph modeling: contrastive learning (BYOL, InfoNCE-style), masked modeling / masked

autoencoders (AudioMAE), **graph neural networks for physiological time-series** (K-NN + temporal edges, masked-node reconstruction), few-shot transfer, cross-subject / cross-dataset generalization.

Generative / multimodal language modeling: LLaMA-2 fine-tuning, LoRA / parameter-efficient adaptation, Q-Former projection, GRPO post-training (VERL), **DAPO preference optimization**, instruction tuning, chain-of-thought supervision, sensor-grounded QA.

ML / stats: PyTorch, transformers, audio spectrogram transformer, time-series modeling, mixed-effects models, statistical signal processing, knowledge distillation + quantization, hyperparameter tuning.

Applied health AI: wearable data pipelines, user studies (n=42), commercial app integration (EQUA), multimodal infant-behavior framework, deployment-ready model compression for edge sensing (UniT-style pruning).

TinyML / on-device: unstructured inference-time pruning for MCUs (UniT), adaptive single-microphone acoustic sensing (EfficientMic), quantization-aware deployment, model compression for edge.

Large-scale training & infra: PyTorch, HuggingFace, DDP on $4 \times$ A100 (80 GB), Slurm HPC, mixed precision (fp16), vLLM, Docker, Weights & Biases, LangChain, **quantization + knowledge distillation**.

Programming: Python, MATLAB, R (working), CUDA (basic).

EXPERIENCE & SERVICE

Graduate Research Assistant

Aug 2022 – Present

Worcester Polytechnic Institute — Advisor: Dr. Bashima Islam

- Lead or co-author on peer-reviewed work at **ACM IMWUT, IEEE BSN, IEEE/ACM CHASE, EMNLP, Interspeech, ICLR, EWSN, UbiComp, and BuildSys**.
- **Reviewer:** IEEE BSN, ACM IMWUT, IEEE Transactions on Affective Computing.
- **Co-organizer:** IEEE ASRU 2025 satellite workshop on AI for Children's Speech and Language.

Graduate Teaching Assistant

Aug 2023 – May 2024

Worcester Polytechnic Institute

- On-Device Deep Learning; Discrete-Time Signal & System Analysis; Communications & Networks.

Assistant Engineer

Aug 2017 – Aug 2022

Electricity Generation Company of Bangladesh — pre-PhD industry experience