KEHAN QI

Ph.D. Student | kehan.qi@stonybrook.edu | (934)263-2748
Personal Page | Google Scholar | Github

Education

Stony Brook University Stony Brook, NY, US
Ph.D. Student in Biomedical Informatics *Aug 2024 - Aug 2028 (expected)*

University of Chinese Academy of Sciences Shenzhen, China
M. Eng. in Computer Technology *Sep 2018 - Jun 2021*

Zhejiang University Hangzhou, China
B. Eng. in Measurement Control Technology and Instruments *Sep 2013 - Jul 2017*

Work Experience

Stony Brook University Stony Brook, NY, US
Research Assistant *May 2025 - present*

- **Flow-based MRI Reconstruction**: Derived that unrolled networks are discretized conditional probability flows; proposed ODE-consistent training objective for unrolled networks; improved PSNR/SSIM by 0.85/0.0092 vs E2E-VarNet on Brainweb dataset 1810 slices. [PDF]
- **MLLM-based Pathology Image Processing**: Built slide-level token compression & query-aware routing for token reduction; reduced GPU memory by 38.9% and TFLOPs by 81.9% with 0.015 accuracy drop on WSI-Bench. [PDF]

Stori Hangzhou, China
Data Engineer (MLE focus) *Apr 2023 - Jul 2024*

- **Real-time Data Pipeline Infrastructure**: Architected and built real-time data pipelines using AWS, and achieved ∼10s latency.
- **Low-latency ML Inference System for Risk Control**: Owned end-to-end design, implementation, and deployment of a real-time ML inference system for transaction-level risk control using AWS, achieved 100ms average latency.
- **Data Infrastructure Monitoring System**: Designed and built the data monitoring system using AWS, achieved ∼1min latency.
- **Query API Integration**: Designed and built data ingestion and query API, achieved ∼10s ingestion latency and ∼100ms query latency, and supported ∼10 QPS.
- **Team Leadership and Standards**: Established internal coding and deployment standards, CI/CD pipeline, and AWS CDK infrastructure templates. Mentored two junior engineers and led weekly sprint planning and code reviews.

Amazon Beijing, China
Software Development Engineer (Applied ML System) *Aug 2021 - Feb 2023*

- **Applied ML System Engineering**: Designed and implemented an automated pipeline for scheduled ingestion of ML model prediction data using AWS, ingesting ∼10M items in 3 hours.
- **Impact Analysis via Distributed Processing**: Built large-scale data processing pipelines to evaluate the financial impact of ML predictions using Spark. Analyzed ∼10B records within 20 minutes.
- **Future Fee Prediction System**: Generated daily future fee estimation reports by combining pre-launch ML predictions with fee rules; batch-processed 1.5B records within 24-hour SLA.

Tencent Shenzhen, China
Research Intern *Jun 2020 – Sep 2020*

- **Medical Image Processing Research**: Conducted research on CT image processing methods for registration and quality assessment.
- **Registration Quality Assessment**: Developed a landmark-based neural network for evaluating registration quality of medical images.

Selected Papers

- **Kehan Qi**, Saumya Gupta, Xiaoling Hu, Qingqiao Hu, Weimin Lyu, and Chao Chen. "Unrolled Networks Are Conditional Probability Flow ODEs in MRI Reconstruction". arXiv preprint arXiv:2512.03020.
- Qingqiao Hu, Weimin Lyu, Meilong Xu, **Kehan Qi**, Xiaoling Hu, and Chao Chen. "LoC-Path: Learning to Compress for Pathology Multimodal Large Language Models." arXiv preprint arXiv:2512.05391.

Skills

- **Languages**: Python, SQL, Java
- **ML/Generative Models**: PyTorch, Diffusion Models, flow matching, Diffusers, Transformers
- **Systems**: AWS (SageMaker, Lambda, Kinesis, DynamoDB, S3, Glue, EMR, VPC), CI/CD, CDK
- **Evaluation**: PSNR/SSIM, ablations, significance tests, p95/p99 latency, TFLOPs
- **Data**: Spark, Flink, Hive, MySQL, NoSQL