XIAOYU (SIMON) SONG

+1(412) 608-6246 ◇ +86 15021173913 ◇ Shanghai
xiaoyuso@alumni.cmu.edu ◇ LinkedIn ◇ GitHub

EDUCATION

Master of Information System Management Aug. 2024 - Dec. 2025
Carnegie Mellon University, Heinz College *Pittsburgh, PA*

BE in Data Science Sep. 2020 - Jun. 2024
Shanghai University, School of Communication and Information Engineering *Shanghai, China*

SUMMARY

Software engineer specializing in distributed memory systems and KV cache infrastructure for large language model inference with research experience and interest in multimodal healthcare AI application.

SKILLS

Programming & Systems C/C++, Python, Linux, Git, Distributed Systems, KV Cache
ML Frameworks PyTorch, TensorFlow, vLLM, SGLang

EXPERIENCE

Software Engineer Mar. 2026 – Present
Huawei Technologies Co., Ltd. *Shanghai, China*

- Developed core features for the open-source Ascend MemCache and MemFabric Hybrid projects using C++, Python, and Linux, focusing on distributed memory management and high-performance data transfer.
- Developed KV cache storage and transfer solutions for LLM inference frameworks, highlighted by integrating Ascend MemCache with SGLang HiCache as a distributed L3 backend for KV cache offloading and reuse.

Research Intern Jan. 2025 – Present
Carnegie Mellon University, Safe AI Lab (Advisor: William Han) *Pittsburgh, PA*

- Currently exploring reasoning and reinforcement learning (RL) techniques to strengthen the ECG-LLM's ability to perform clinical interpretation, evidence retrieval, and multi-step diagnostic reasoning.
- Extended the ECG-Byte model with retrieval-augmented capability to enhance multimodal representation learning and improve end-to-end performance in ECG–text large language modeling.
- Integrated statistical and pathological feature retrieval into prompting pipelines, improving BLEU-4, ROUGE-L, METEOR, BERTScore-F1, and accuracy by 66.7%, 13.1%, 18.4%, 1.3%, and 89.3%, respectively.
- Conducted extensive ablation studies on retrieval strategies, report complexity, and prompt structure to identify key design factors influencing model interpretability and reasoning performance.

Algorithm Engineer Intern May 2025 – Aug. 2025
Anker Innovations *Shanghai, China*

- Designed and implemented an LSTM-based deep learning model for real-time heart rate estimation from multimodal PPG+IMU wearable signals, achieving an average prediction error of 3 BPM.
- Engineered an object-oriented experimentation framework supporting modular configuration, automated training, evaluation, and performance logging for large-scale model iteration.
- Applied model pruning and quantization techniques to reduce computational cost (50% FLOPs, 85% memory) with minimal performance degradation (+0.7 BPM).

Algorithm Engineer Intern Dec. 2023 – Apr. 2024
Shanghai Intelligent and Connected Vehicle R&D Center Co., Ltd. *Shanghai, China*

- Collaborated with enterprise clients to define customized evaluation metrics (e.g., Relative Delay) for multi-source traffic congestion detection in underground parking systems.
- Integrated Nvidia DeepStream with YOLOv5 for real-time object tracking and vehicle trajectory analysis across camera networks.
- Implemented hash table–based caching for vehicle–frame mapping, achieving a 6.6% improvement in processing efficiency.

Intelligence Analyst Intern Jun. 2023 – Aug. 2023
Meituan & Dianping *Shanghai, China*

- Applied XGBoost, unsupervised clustering, and funnel analysis to construct behavioral profiles for user segmentation and fraud detection.
- Built interpretable risk modeling pipelines that generated over 10 high-quality detection cases daily, directly supporting real-time risk control operations.
- Automated large-scale data preprocessing and anomaly visualization workflows using Python, improving analytical throughput and consistency.

Research Intern Nov. 2022 – Sep. 2023
Shanghai Jiao Tong University (Advisors: Prof. Jun Wu and Hansong Xu) *Shanghai, China*

- Applied Instant-NGP NeRF to reconstruct 3D building models from multi-view imagery.
- Generated 3D urban environments by converting 2D maps using Unity and CityGen3D.
- Built an interactive digital twin by integrating NeRF models with simulated city maps.
- Combined YOLO and SAM for object detection and segmentation to support robotic manipulation.

Asset Management Analyst Intern Jul. 2022 – Aug. 2022
Industrial Securities Futures *Shanghai, China*

- Designed Python- and Excel-based quantitative analysis tools for evaluating portfolio risk metrics, including Sharpe ratio, maximum drawdown, and return distribution.
- Automated document parsing and database ingestion workflows, reducing manual processing time by 90%.
- Generated comprehensive multi-dimensional performance reports to inform strategic asset allocation and portfolio optimization.

SELECTED PUBLICATIONS AND PREPRINTS

- **Xiaoyu Song***, William Han*, Tony Chen, Chaojing Duan, Michael Rosenberg, Emerson Liu, Ding Zhao. Retrieval-Augmented Generation for Electrocardiogram-Language Models. ArXiv.org, 30 Sep. 2025, arxiv.org/abs/2510.00261.
- William Han, Chaojing Duan, Zhepeng Cen, Yihang Yao, **Xiaoyu Song**, Atharva Mhaskar, Dylan Leong, Michael Rosenberg, Emerson Liu, Ding Zhao. Signal, Image, or Symbolic: Exploring the Best Input Representation for Electrocardiogram-Language Models Through a Unified Framework. ArXiv.org, 24 May 2025, arxiv.org/abs/2505.18847.
- **Xiaoyu Song**, Hansong Xu, Xing Liu, Fan Liang. An Integrated Instant NeRF and Simulation Based 3D Reconstruction for Immersive City Twin. In: Park, J.S., Yang, L.T., Pan, Y., Park, J.J. (eds) *Advances in Computer Science and Ubiquitous Computing (CUTECSA 2023)*, Lecture Notes in Electrical Engineering, vol. 1190. Springer, Singapore, 2024, https://doi.org/10.1007/978-981-97-2447-5_8
- **Xiaoyu Song**. Prediction of People's Abnormal Behaviors Based on Machine Learning Algorithms. *Proc. 2022 Int. Conf. on Machine Learning and Intelligent Systems Engineering (MLISE)*, Guangzhou, China, 2022, pp. 406–409, https://doi.org/10.1109/MLISE57402.2022.00087