

Ziliang (Johnson) Zhang

 zzhan357@ucr.edu  858-699-1996  Website  LinkedIn  Google Scholar  Riverside, CA

Education

University of California, Riverside

09/2022 ~ (Exp.) 12/2026

Doctor of Philosophy, Electrical and Computer Engineering

Research focus: Real-time AI systems, specialized in recommender systems and AI wearables

University of California, Riverside

09/2019 ~ 03/2021

Master of Science, Computer Engineering

Study area: real-time systems, computer architecture, on-device machine learning

University of California, San Diego

08/2015 ~ 03/2019

Bachelor of Science, Cognitive Science: Human-Computer Interaction

Study area: Artificial Intelligence, Parallel Algorithms, Data Analytics, Software Development, UI/UX

Work Experience

AI/ML Engineer Intern

LinkedIn

06/2026 ~ 09/2026

- Researched LinkedIn Generation Recommendation (GR) with 1.7B monthly traffic in CoreAI team.
- Addressed zero-history cold start with publication and achieved 34% hit rate and NDCG increase.

GPU Coder Internship

MathWorks

06/2025 ~ 09/2025

- Developed GPU optimization in ML compiler/IR to speed up the CUDA production code by 17x.
- Researched AI algorithm on CUDA codegen from MATLAB and increased 76% GPU Utilization.

Full-Stack Software Developer

Verizon

07/2021 ~ 09/2022

- Developed micro-service APIs with Spring Framework, Redis, and Angular for 70M monthly traffic.

Highlight Publications

Zhang et al. *ZeroRec: Harnessing Uncertainty-guided Principles for Zero-History Recommendation*

- Calibrated world knowledge with bayesian in post-trained Generative Rec and improves 33% HR.

Z. Zhang, H. Kim, C. Liu. *RELAXR: Real-time Execution Framework for Latency-aware AI-native Extended Reality*. In Submission

- Prototyped agentic AI glasses and improved real-time latencies in both XR and AI pipelines by 56%.

Z. Zhang, C. Liu, H. Kim. *MATER: Mutually Aware Framework for Teleoperated-robot with Extended Reality*. ACM SIGMETRICS 2026 (SIGMETRICS'26)

- Overhauled XR teleoperation system with ML approximation that shortens completion time by 48%.

Y. Chen*, **Z. Zhang*** (Co-first), H. Kim, C. Liu. *REFiNE: Reward and Energy Fairness for Robust Multi-Agent Coordination in RL-driven Robots*. In Submission

- Achieved fairness in Multi-agent Reinforcement Learning scenarios that prolongs runtime by 22%.

A. Bukhari, B Mamo, M. S. Hossain, **Z. Zhang**, M. Karimi, D. Enright, P. Manosalva, H. Kim. *Low-Cost Sensing and Classification for Early Stress and Disease Detection in Avocado Plants* (IEEE IoT-J'25)

- Proposed disease detection methods using LSTM and DNN algorithm to achieve 86% accuracy.

Z. Zhang, Z. Li, H. Kim, C. Liu. *BOXR: Body and head motion Optimization framework for eXtended Reality*. IEEE Real-Time Systems Symposium (RTSS'24)

- Discovered the Camera-to-Display latency and optimized XR framework to reduce 63% latency.

Z. Zhang, C. Liu, H. Kim. *MII: A Multifaceted Framework for Intermittence-aware Inference and Scheduling*. ACM International Conference on Embedded Software (EMSOFT'24)

- Delivered layer-wise checkpointing and real-time scheduling to increase 39% successful jobs.

Z. Zhang, C. Liu, H. Kim. *Toward a Predictive eXtended Reality Teleoperation System with Duo-Virtual Spaces*. Southern California Robotics Symposium (SCR'24)

- Proposed XR-robot teleoperation predictive display that reduces teleoperation error by 73%.

Y. Zhuy, H. Alrashid, S. Bai, C. Zhang, **Z. Zhang**, Z. Qu, R. Y. Ali, A. Magdy. *On the Ecosystem of High-Definition (HD) Maps*. IEEE 40th ICDEW (ICDEW'24)

- Analyzed over 53 state-of-the-art approaches in HD Maps construction, application, and evaluation.

Project Experience

Predictive Display model and GPU code acceleration @ **MathWorks** (MATLAB, CUDA, Transformer)

- Proposed predictive display transformer model and achieved 6.6X speed up.

Real-Time Linux Kernel @ **RTEN Lab, UCR** (C, Scheduler, Real-time Clock, Linux Kernel)

- Developed IOCTL, LKM, High-resolution Timer, and TCB for real-time support and maintained high QoS and Timeliness Correctness of 1k+ concurrent tasks.

Academic Service

Reviewer @ Transactions on Embedded Computing Systems (TECS)	2025
Reviewer @ Journal of Systems Architecture (JSA)	2025
Secondary Reviewer @ Association for Computational Linguistics (ACL)	2025
External Reviewer @ IEEE Real-Time Systems Symposium (RTSS)	2024, 2025
External Reviewer @ ACM International Conference on Embedded Software (EMSOFT)	2024
External Reviewer @ Euromicro Conference on Real-Time Systems (ECRTS)	2024
Teaching Assistant @ Embedded Systems (EE128)	2023, 2024, 2025
Teaching Assistant @ Real-Time Embedded Systems (EE255)	2024
Teaching Assistant @ Computing Thinking (EDS124BR)	2019

Honors and Awards

Outstanding Teaching Award, UCR	2025
Conference Travel Grant, UCR	2024
Dean's Distinguished Fellowship Award, UCR	2022
Provost Honor, UCSD	2016, 2018

Technical Skills

Machine Learning: Tensorflow, PyTorch, ONNX, RecSys, RLHF, Distillation, RAG, Retrieval, Search;

GPU Programming: CUDA, CUDNN, GPU codegen/compiler, Kernel Scheduler, CUDA Graph;

Distributed Systems: AWS (EC2, ECS, DynamoDB, Lambda), LLM serving, Collaborative Inference;

On-device ML: NAS, FreeRTOS, LKM, Real-time Scheduling, ROS2, Networking, efficient LLM;

Programming Language: Python, Java 8/11, HTML5/CSS3/Javascript, Typescript, R;

CI/CD: Perforce, Git, JIRA, Scrum, Kibana, Kubernetes, Docker, Postman;