Qinsi Wang

Durham, North Carolina, United States.| +1 919-672-0619 | qinsi.wang@duke.edu | https://wangqinsi1.github.io

Research Interests

Data-Centric LLM/VLM Post-Training; Agent Self-Improvement; Efficient LLM/VLM Inference

Education

Duke University — Durham, North Carolina, United States
Ph.D in Electrical and Computer Engineering. Advisor: Prof. Yiran Chen. — *Aug. 2024 – Present*

Huazhong University of Science and Technology — Wuhan, Hubei, China
B.S. in Electronic Science and Technology Major — *Sep. 2018 – June 2022*

Work Experience

Meta — Burlingame, California, United States
Research Intern. Manager: Qian Qi, Wei Wen. — *May. 2026 – Now*

Key Achievements: *Developed egocentric AI assistant to enhance the intelligence and contextual awareness of Meta AI glasses. Investigated how embodied assistants perceive, reason about, and interact with the physical world.*

Adobe — San Jose, California, United States
Research Intern. Managers: Wentian Zhao and Daniel Miranda — *May 2025 – Oct. 2025*

Key Achievements: *Developed a self-improving agent system to enhance autonomous learning and performance. Published three first-author papers and received Adobe's Gift Fund in recognition of outstanding intern performance.*

First-Author Publications

Post-Training / Agentic Self-Improvement

- **Qinsi Wang**, Jing Shi, Huazheng Wang, Kun Wan, Yiran Wu, Bo Liu, Qingyun Wu, Hai Helen Li, Yiran Chen, Handong Zhao, Wentian Zhao, "From RLVR to RLSVR: Breaking the Verifiability Bottleneck Beyond Math and Code via Self-Verifiable Rewards", Third Conference on Language Modeling (***COLM***) 2026.
- **Qinsi Wang**, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, Wentian Zhao, "Vision-Zero: Scalable VLM Self-Improvement via Strategic Gamified Self-Play", International Conference on Learning Representations (***ICLR***), 2026.
- **Qinsi Wang**, Jinghan Ke, Hancheng Ye, Yueqian Lin, Yuzhe Fu, Jianyi Zhang, Kurt Keutzer, Chenfeng Xu, Yiran Chen, "Angles Don't Lie: Unlocking Training-Efficient RL Through the Model's Own Signals", Neural Information Processing Systems (***NeurIPS Spotlight***), 2025.

Multimodal Reasoning

- **Qinsi Wang**, Wei Wen, Chong Zhou, Saksham Suri, Zechun Liu, Lemeng Wu, Hai Helen Li, Yiran Chen, Raghuraman Krishnamoorthi, Qi Qian, "InteractiveWorld: A Verifiable Human-Centric Environment for Building Better Interactive MLLM Assistants", submitted to ICLR 2027, arXiv:2610.33966.
- **Qinsi Wang**, Jing Shi, Kun Wan, Handong Zhao, Hancheng Ye, Zishan Shao, Jinghan Ke, Yudong Liu, Daniel Miranda, Purvak Lapsiya, Yiran Chen, Wentian Zhao, "Seeing is Solving: Unlocking Efficient Multimodal RL via View Alignment", International Conference on Machine Learning (***ICML***) 2026.
- **Qinsi Wang***, Hancheng Ye*, Jinhee Kim*, Jinghan Ke, Yifei Wang, Martin Kuo, Zishan Shao, Dongting Li, Yueqian Lin, Ting Jiang, Chiyue Wei, Qi Qian, Wei Wen, Helen Li, Yiran Chen, "T2S-Bench & Structure-of-Thought: Benchmarking and Prompting Comprehensive Text-to-Structure Reasoning", arXiv:2603.03790.

Efficient LLMs / VLMs

- **Qinsi Wang***, Jinghan Ke*, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, Chenfeng Xu, "Dobi-SVD: Differentiable SVD for LLM Compression and Some New Perspectives", International Conference on Learning Representations (***ICLR***), 2025.

- **Qinsi Wang**, Jianyi Zhang, Jianyang Gu, Hancheng Ye, Saeed Vahidian, Yiran Chen, "CoreInfer: Accelerating Large Language Model Inference with Semantics-Inspired Adaptive Sparse Activation", International Conference on Machine Learning (***ICML***) 2026.
- **Qinsi Wang**, Hancheng Ye, Ming-Yu Chung, Yueqian Lin, Martin Kuo, Mingyuan Ma, Jianyi Zhang, Yiran Chen, "CoreMatching: A Co-adaptive Sparse Inference Framework with Token and Neuron Pruning for Comprehensive Acceleration of Vision-Language Models", International Conference on Machine Learning (***ICML***) 2025.
- **Qinsi Wang***, Jinghan Ke*, Zhi Liang, Sihai Zhang, "MathNAS: If Blocks Have a Role in Mathematical Architecture Design", Neural Information Processing Systems (***NeurIPS***), 2023.
- **Qinsi Wang**, Sihai Zhang, "DGL: Device Generic Latency model for Neural Architecture Search on Mobile Devices.", IEEE Transactions on Mobile Computing (***IEEE TMC***), 2023.

Other Publications

Conference & Journal Papers

- Yueqian Lin, **Qinsi Wang**, Hancheng Ye, Yuzhe Fu, Hai Li, Yiran Chen, "Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding", Conference on Computer Vision and Pattern Recognition (***CVPR***), 2026.
- Yudong Liu, Jingwei Sun, Yueqian Lin, Jingyang Zhang, Ming Yin, **Qinsi Wang**, Jianyi Zhang, Hai Helen Li, Yiran Chen, "Keyframe-oriented vision token pruning: Enhancing efficiency of large vision language models on long-form video processing", International Conference on Computer Vision (***ICCV***), 2025.
- Zishan Shao, Yixiao Wang, **Qinsi Wang**, Ting Jiang, Zhixu Du, Hancheng Ye, Danyang Zhuo, Yiran Chen, Hai Li, "FlashSVD: Memory-Efficient Inference with Streaming for Low-Rank Models", the Association for the Advancement of Artificial Intelligence (***AAAI***), 2026.
- Hancheng Ye, Zhengqi Gao, Mingyuan Ma, **Qinsi Wang**, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, Yiran Chen, "Kvcomm: Online cross-context kv-cache communication for efficient llm-based multi-agent systems", Neural Information Processing Systems (***NeurIPS***), 2025.
- Chiyue Wei, Cong Guo, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu, **Qinsi Wang**, Changchun Zhou, Hai Li, Yiran Chen, "Focus: A Streaming Concentration Architecture for Efficient Vision-Language Models", High-Performance Computer Architecture (***HPCA Best Paper Candidate***), 2026.
- Yueqian Lin, **Qinsi Wang**, Yudong Liu, Hancheng Ye, Hai Li, Yiran Chen, "RAISE: Resolving Ambiguity in Audio Understanding with Imagination and Selective Extraction", ***Interspeech***, 2026.
- Yuzhe Fu, Hancheng Ye, Cong Guo, Junyao Zhang, **Qinsi Wang**, Yueqian Lin, Changchun Zhou, Hai Li, Yiran Chen, "FlashFPS: Efficient Farthest Point Sampling for Large-Scale Point Clouds via Pruning and Caching", Design Automation Conference (***DAC***), 2026.
- Yueqian Lin, Zhengmian Hu, Jayakumar Subramanian, **Qinsi Wang**, Nikos Vlassis, Hai Li, Yiran Chen, "AsyncVoice Agent: Real-Time Explanation for LLM Planning and Reasoning", IEEE Automatic Speech Recognition and Understanding Workshop (***ASRU***) Demo Track, 2025.
- Yixiao Wang, Ting Jiang, Zishan Shao, Hancheng Ye, Jingwei Sun, Mingyuan Ma, **Qinsi Wang**, Jianyi Zhang, Yiran Chen, Hai Helen Li, "ZEUS: Accelerating Diffusion Models with Only Second-Order Predictor", ACM International Conference on Multimedia (***ACMMM***), 2026.

Preprints

- Martin Kuo, Jianyi Zhang, Aolin Ding, **Qinsi Wang**, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, Yiran Chen, "H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking", arXiv:2502.12893.
- Yudong Liu, Yuan Li, Zijia Tang, Yuxi Zheng, Yueqian Lin, **Qinsi Wang**, Yi Li, Shuangjun Liu, Shuai Zhang, Taotao Jing, Dashan Gao, Ning Bi, Jingwei Sun, Yiran Chen, Hai Li, "Latent Bridge: Feature Delta Prediction for Efficient Dual-System Vision-Language-Action Model Inference", arXiv:2605.02739.
- Yifei Wang, Hancheng Ye, Yechen Xu, Cong Guo, Chiyue Wei, **Qinsi Wang**, Dongting Li, Tingjun Chen, Hai Helen Li, Danyang Zhuo, Yiran Chen, "MARS: Efficient, Adaptive Co-Scheduling for Heterogeneous Agentic Systems", arXiv:2604.26963.
- Yueqian Lin, Zhengmian Hu, **Qinsi Wang**, Yudong Liu, Hengfan Zhang, Jayakumar Subramanian, Nikos Vlassis, Hai Li, Yiran Chen, "Voice Evaluation of Reasoning Ability: Diagnosing the Modality-Induced Performance Gap", arXiv:2509.26542.

Honors & Awards

- Excellent Graduate of Huazhong University of Science and Technology. 2022
- **National Scholarship (top 0.2%, highest scholarship from Ministry of Education of China).** 2020
- Outstanding Student Scholarship (top 5%). 2019 & 2020
- Chinese Mathematics Competitions, First Prize (top 8%). 2019
- Mathematical Contest In Modeling(MCM/ICM), Honorable Mention (top 30%). |*Team leader.* 2020
- China Undergraduate Mathematical Contest in Modeling(CUMCM), Second Prize (top 10%).|*Team leader.* 2020