Xinrui Zhong

xinruizh@andrew.cmu.edu

zxr-creator.github.io | Google Scholar | GitHub | LinkedIn

RESEARCH INTERESTS

Efficient Model Architectures, Machine Learning Systems, and Recursive Self-Improvement (AI agents that build and optimize ML systems).

EDUCATION

- **Rice University** *Aug 2024 – Dec 2025*
  Master of Computer Science Houston, USA
  - GPA: 4.00/4.00
- **Sichuan University** *Sep 2020 – Jun 2024*
  Bachelor of Engineering in Computer Science and Technology Chengdu, China
  - GPA: 3.66/4.00 | Major GPA: 3.72/4.00

PUBLICATIONS

2026 **ServerBench: How can your coding agents design specialized LLM serving systems?**
Xinrui Zhong, Zhiming Wang, Yongqi Liu, Zhaohui Wang, Shanglin Yang, Zikun Li, Zhuoming Chen, Beidi Chen
Submitted to NAACL 2027 [arXiv] [Code] [Blog]

2026 **CompressionBench: How can your coding agents compress LLMs?**
Xinrui Zhong, Zhifeng Li, Zhiming Wang, Yongqi Liu, Shanglin Yang, Yang Zhou, Zikun Li, Zhuoming Chen, Beidi Chen
Submitted to MLSys 2027 [arXiv] [Code] [Blog]

2026 **RISA: Recoverable Sparse Attention for Block-Diffusion Language Models**
Xinrui Zhong, Fanjiang Ye, Yuke Wang
Submitted to ICLR 2027 [arXiv]

2026 **Vortex: Efficient and Programmable Sparse Attention Serving for AI Agents**
Zhuoming Chen, **Xinrui Zhong**, Qilong Feng, Ranajoy Sadhukhan, Yang Zhou, Michael Qizhe Shieh, Zhihao Jia, Beidi Chen
NeurIPS 2026 [arXiv] [Code] [Website]

2026 **SparSP: Exploiting Communication Sparsity for Sequence-Parallel Video DiTs**
Desen Sun, **Xinrui Zhong**, Yuke Wang, Sihang Liu
Submitted to EuroSys 2027 [arXiv]

2026 **GENSERVE: Efficient Co-Serving of Heterogeneous Diffusion Model Workloads**
Fanjiang Ye, Zhangke Li, **Xinrui Zhong**, Ethan Ma, Russell Chen, Kaijian Wang, Jingwei Zuo, Desen Sun, Ye Cao, Triston Cao, Myungjin Lee, Arvind Krishnamurthy, Yuke Wang
Submitted to MLSys 2027 [arXiv]

2026 **R^2-dLLM: Accelerating Diffusion Large Language Models via Spatio-Temporal Redundancy Reduction**
Zhenbang Du, Kejing Xia, **Xinrui Zhong**, Yonggan Fu, Nicolai Oswald, Binfei Ji, Brucek Khailany, Pavlo Molchanov, Yingyan (Celine) Lin
Submitted to NAACL 2027 [arXiv] [Code]

2025 **An Efficient and Adaptive Watermark Detection System with Tile-based Error Correction**
Xinrui Zhong, Xinze Feng, Jingwei Zuo, Fanjiang Ye, Yi Mu, Junfeng Guo, Heng Huang, Myungjin Lee, Yuke Wang
Submitted to MLSys 2027 [arXiv]

RESEARCH EXPERIENCE

- **Carnegie Mellon University (InfiniAI Lab)** *Feb 2026 – Present*
  Research Intern, advised by Prof. Beidi Chen Pittsburgh, USA (Hybrid)
  - Co-developed **Vortex**, a programmable sparse-attention serving framework with a Python-embedded vFlow DSL and a page-centric vTensor abstraction for composing dynamic KV-cache routing over paged and ragged layouts.
  - Integrated Vortex with production serving backends and optimized sparse execution through workload-aware planning, operator fusion, and radix top-k selection, achieving up to 3.46× throughput over full attention while preserving accuracy.
  - Led **ServerBench**, a benchmark in which coding agents build complete, specialized LLM serving systems, graded by stateful session replay and independent correctness verification. Evaluated 11 agent models across three deployments using latency-qualified token value per dollar of GPU cost; the best agent-built system reached 4.87× the vLLM reference on Gemma E4B.
  - Led **CompressionBench**, a benchmark in which coding agents compress pinned BF16 LLM checkpoints under 8–32 GiB storage budgets with any mix of quantization, pruning, low-rank factorization, and distillation; a hidden evaluator checks package size, GPU memory, and fidelity once, after the deadline.

- **Rice University (Yuke's Lab)** *Jan 2025 – Present*
  Research Intern, advised by Prof. Yuke Wang Houston, USA
  - Developed **RISA**, recoverable sparse attention for block-diffusion LMs: newly generated history is sorted into exact priority and recovery tiers, and seed queries build one bounded read plan reused by every denoising step of a block. On DreamReasoner-8B, attention is 2.67× faster than dense at 32K and 5.53× at 128K.
  - Developed **QRMark**, a tile-based watermark detection method with Reed–Solomon error correction that exploits redundancy across tiles to balance detection efficiency, scalability, and robustness; the end-to-end pipeline is up to 2.42× faster than existing baselines on Stable Signature and AquaLoRA.
  - Co-developed **GenServe**, an SLO-aware system for co-serving heterogeneous diffusion workloads using patch packing, spatial partitioning, preemption, and elastic sequence parallelism to adapt GPU allocation.

- **Rice University (DATA Lab)** *Sep 2024 – Dec 2024*
  Research Intern, advised by Prof. Xia Hu Houston, USA
  - Studied image-watermark detection pipelines and the trade-off between detection accuracy and efficiency, which became the starting point for QRMark.
  - Contributed to the training and evaluation pipeline of LTSM-Bundle, including TDengine-backed data loading, and validated models across diverse time-series datasets. Code: github.com/datamllab/ltsm.

SELECTED PROJECTS

- **Prompt Enhancement Framework for Stable Diffusion** *Jan 2025 – Apr 2025*
  Course project, COMP 646: Deep Learning for Vision and Language, advised by Prof. Vicente Ordonez
  - Fine-tuned Qwen2.5 for prompt enhancement and Stable Diffusion 2.1 for image generation; the end-to-end pipeline improved prompt alignment and image quality, with higher IS and CLIP scores than the baseline.

ACADEMIC SERVICE

- Artifact Evaluation Committee Member: SOSP 2026, MLSys 2026, PPoPP 2026