Runkai Tao

732-668-6861 | rt572@physics.rutgers.edu | my website | my github

Education

Rutgers University — New Brunswick, NJ, US
Ph.D. Student in Physics, Advisor: Greg Moore, Heeyeon Kim — *Aug. 2019 – present*

Fudan University — Shanghai, China
Bachelor of Science in Physics, Advisor: Satoshi Nawata — *Sep. 2014 – July 2019*

Research Experience

Amazon AWS — Fall 2025
Applied Scientist Intern

Efficient and Intelligent Computing Lab, Georgia Tech — Spring 2025
Intern

- Distributed system design for GNN

Department of Physics, Rutgers University — July 2019 – present
Research Assistant, Advisor: Greg Moore, Heeyeon Kim

- Research on various aspects on supersymmetric gauge theories and string theory
- Research on non-chiral CFT and VOA

Department of Physics, Fudan University — June 2018 – May 2019
Research Assistant, Advisor: Satoshi Nawata

- Research on representation theory of knots
- Research on the N=(0,2) Landau-Ginzburg model

Projects

SGLang:

- Hybrid KV cache for LLaMA-4:

  LLaMA-4 uses **local attention** in three-quarters of its layers. To balance KV-cache memory, I split the cache into global and local ones and rewrote the **eviction policy** for chunked prefill and decoding. Tuning their relative sizes of the two caches yields $\approx 10\%$ higher throughput on long-context inference and supports context lengths up to three times compared to the baseline.

  Related PR #6653, #6657, #5575;

  Related techniques: **CUDA Graph**, **KV-cache**.

- Data Parallelism Attention for DeepSeek Models:

  To minimize memory usage in DeepSeek models, we employ **data parallelism** (DP) within the self-attention layer. When the DP size for self-attention is smaller than the **expert-parallel** (EP) size in the MoE layers, we use **tensor parallelism** (TP) inside each DP batch to reduce latency. We apply **reduce-scatter** operation to distribute tokens before the MOE layer, and an **all-gather** afterward to restore the tensor.

  Related PR: #4770;

  Related techniques: **NCCL collectives**, **Parallel techniques for LLMs**.

Distributed system design for GNN:

- RNS-GCN, MixGCN:

  To cut memory usage and accelerate both training and inference, I merge the feature-update and **all-to-all** communication into a single PyTorch layer, implemented a custom backward pass, and rewrote the **Triton kernel** for better performance.

  Related techniques: **PyTorch**, **Triton**, **GNN**.

Publications and Preprints

- Path Integral Derivations of K-Theoretic Donaldson Invariants, (with Heeyeon Kim, Jan Manschot, Gregory W. Moore and Xinyu Zhang,) (to appear).
- Argyres-Douglas Theories, Macdonald Indices And Arc Space Of Zhu Algebra, (with Anindya Banerjee, Singh, Ranveer Kumar,) arXiv preprint arXiv:2507.06294 (2025)
- MixGCN: Scalable GCN Training by Mixture of Parallelism and Mixture of Accelerators, (with Wan Cheng, Zheng Du, Yang Katie Zhao, and Yingyan Celine Lin,) arXiv preprint arXiv:2501.01951 (2025).
- Rationality of Lorentzian Lattice CFTs and the Associated Modular Tensor Category, (with Singh, Ranveer Kumar, Madhav Sinha,) arXiv preprint arXiv:2408.02744 (2024).
- Fudan Lectures on 2d Conformal Field Theory, (with Nawata Satoshi and Daisuke Yokoyama,) arXiv preprint arXiv:2208.05180 (2022).
- Fudan Lectures on String Theory, (with Nawata Satoshi and Daisuke Yokoyama,) arXiv preprint arXiv:2208.05179 (2022).
- New Conformal Field Theory from N=(0,2) Landau-Ginzburg Model, (with Guo Jirui, Satoshi Nawata and Hao Derrick Zhang,) Physical Review D 101, no. 4 (2020): 046008.
- Cyclotomic Expansions of HOMFLY-PT Colored by Rectangular Young Diagrams, (with Kameyama Masaya, Satoshi Nawata, and Hao Derrick Zhang,) Letters in Mathematical Physics 110 (2020): 2573-2583.

Teaching Experience

Teaching Assistant at Rutgers University Spring 2025
Physics 203,204: General Physics

Teaching Assistant at Rutgers University Fall 2020
Physics 229: Analytical physics II Laboratory

Teaching Assistant at Rutgers University Summer 2020
Physics 205,206: General Physics Laboratory

Teaching Assistant at Rutgers University Spring 2020
Physics 382 , Lecturer: Weida Wu

Teaching Assistant at Rutgers University Fall 2019
Physics 381 , Lecturer: Weida Wu

Teaching Assistant at Fudan University Spring 2019
Gauge field theory , Lecturer: Yang Zhou

Teaching Assistant at Fudan University Fall 2018
Particle Physics and String Theory , Lecturer: Satoshi Nawata