

TIANYU (MAX) WANG

✉ tianyu.wang@pitt.edu 🏠 Homepage 🔗 LinkedIn 🎓 Google Scholar

Education

University of Pittsburgh, PA, USA

Aug. 2022 – Present

Ph.D. in Computer Science

(Field: Efficient LLM serving, Multi-agent RL Training, Video generation KV management)

Advisor: Dr. Longfei Shangguan

University of Rochester, Rochester, NY

Aug. 2020 – Dec. 2021

M.S. in Computational Science

The Ohio State University, Columbus, OH

Jan. 2017 – Jul. 2020

B.S. in Computer & Information Science, B.S. in Mathematics

Experience

Hewlett Packard Enterprise Labs

May. 2025 – Present

Research Intern

Mentor: Dr. Gourav Rattihalli and Dr. Aditya Dhakal

- **Energy-Aware Scheduling for Serverless LLM Serving on Shared GPUs:** Developed an energy-aware LLM serving prototype **deployed on HPE GreenLake**, integrating workload dispatch and consolidation across shared GPUs. Reduced energy consumption by **56%** while maintaining SLO attainment on par with state-of-the-art serving systems.

University of Pittsburgh

Aug. 2022 – Present

Research Assistant

Advisor: Dr. Longfei Shangguan

- **KV Cache Compression for Long Video Generation:** Developed an online KV-cache compression runtime integrated into SGLang-Diffusion to manage growing KV histories during long-form video generation. Combined on-the-fly compression with prefix-tree-based cache reuse to avoid redundant KV computation, achieving a **>20x compression ratio** and **>5x lower next-chunk latency**.
- **Accelerating Dynamic Sparse Attention for Long-Context Decoding:** Designed a temporal-locality-aware runtime that speculatively computes attention over predicted top-K KV blocks while overlapping execution with KV selection. Combined this scheduling approach with a custom FA3-style kernel, reducing per-token decoding latency by **over 40%**.
- **Efficient RAG Serving:** Designed a multi-GPU prefix KV-cache management policy for RAG serving, combining workload redistribution with out-of-order query execution to coordinate cache usage and query scheduling. Reduced time to first token (TTFT) by **over 64%**.
- **Multi-Tenant Continual Learning:** Proposed an inference-training co-scheduling framework that uses integer linear programming to dynamically reconfigure GPU MIG slices. Coordinated resource allocation across inference and continual-learning workloads, improving SLO attainment and accuracy by **over 1.4x** and **1.49x**, respectively.
- **RL Training Acceleration for Multi-Agent Systems:** Implemented a dependency-aware rollout policy on top of Slime and OWL, coordinating multi-agent execution around inter-agent dependencies to avoid blocked rollouts during RL training.

Enflame Technology Inc., Shanghai, CN

Jan. 2022 – Aug. 2022

Machine Learning Systems and Frameworks Researcher

Mentor: Dr. Heng Shi

- **Raptor Framework:** Developed a Rust-based scheduling framework for distributed DNN training. Formulated parallelism strategy selection as an integer linear program, enabling automatic, optimization-driven planning of parallel execution for distributed training workloads.
- **MLIR Operator Conversion:** Implemented conversions between custom MLIR dialects and built-in dialects, bridging custom operator representations with the MLIR compilation infrastructure.

Publications

- [1] **T. Wang**, G. Rattihalli, A. Dhakal, L. Shangguan, D. Milojevic, "Energy-Aware Scheduling for Serverless LLM Serving on Shared GPUs," (**ATC 2026**)
- [2] **T. Wang**, G. Rattihalli, A. Dhakal, X. Tang, D. Milojevic, "WAGES: Workload-Aware GPU Sharing System for Energy-Efficient Serverless LLM Serving," (**SC 2025 Workshop**)
- [3] **T. Wang**, S. Li, B. Li, Y. Dai, A. Li, G. Yuan, Y. Ding, Y. Zhang, and X. Tang, "Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration," (**JSA**)
- [4] **T. Wang**, D. Yu, Z. Ren, J. Zhang, L. Shangguan, "ReelCache: An Efficient KV Cache Compression System for Long Video Generation," (**under review at MLSys 2027**)
- [5] **T. Wang**, G. Rattihalli, A. Dhakal, J. Li, Z. Ren, D. Milojevic, L. Shangguan, "Predict, Reuse, and Repair: Accelerating Dynamic Sparse Attention for Long-Context LLM Decoding" (**under review at ICLR 2027**)
- [6] S. Li, Q. Tan, Y. Dai, Z. Kong, **T. Wang**, J. Liu, A. Li, N. Liu, Y. Ding, X. Tang, and G. Yuan, "A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning," (**ICLR 2025**)
- [7] B. Li, Y. Wang, **T. Wang**, L. Eeckhout, J. Yang, A. Jaleel, and X. Tang, "STAR: Sub-Entry Sharing-Aware TLB for Multi-Instance GPU," (**MICRO 2024**)

Selected Honors and Awards

University of Pittsburgh SCI PhD Fellowship

2022–2023

University of Rochester Scholarship

2020–2021

Skills

Programming language: Python, Rust, C++, Java

Machine Learning: Pytorch, vLLM, PytorchFSDP, NCCL, DeepSpeed, MLIR, ONNX

Topics: energy-efficiency, LLM serving, serverless, diffusion, video generation