Xiaowen Zhang

xiaowen5@andrew.cmu.edu | handshaker86.github.io

Education

Carnegie Mellon University Pittsburgh, PA
Exchange Student in Electrical and Computer Engineering *Jan. 2026 – Present*

- **Relevant Coursework**: Intro to Computer Systems (18-213), GenAI (10-423)

Shanghai Jiao Tong University Shanghai, China
B.S. in Electrical and Computer Engineering *Sept. 2023 – June 2027 (Expected)*

- **GPA**: 3.75/4.0 — **Rank**: 15/215
- **Relevant Coursework**: Programming and Elem. Data Structures, Data Structures and Algorithms, Intro to Computer Organization

Research & Project Experience

AI for Science & Machine Learning Systems Shanghai Jiao Tong University
Research Assistant (Advised by Prof. David L.S. Hung & Dr. Fengnian Zhao) *Dec. 2024 – Present*

- Developed **FlowForge**, a compile–execute engine designed to tackle computational and latency bottlenecks in physical flow field prediction.
- Engineered an offline compiler that generates domain-aware static execution schedules and lowers them into memory lookup tables, eliminating runtime causal masking overhead.
- Designed a staged local rollout mechanism that enforces bounded memory access and predictable latency, significantly improving execution efficiency and robustness against input corruptions.
- Prototyped a **speculative decoding** pipeline to accelerate autoregressive rollout, orchestrating a lightweight draft model for rapid candidate generation and a high-fidelity model for parallel verification.
- **Publication**: **X. Zhang**, et al. "FlowForge: A Staged Local Rollout Engine for Flow-Field Prediction". *Under review at a premier machine learning conference.*

Tencent Generative Advertising Recommendation Algorithm Track Online
Algorithm Developer | ***Ranked Top 10%*** *Jun. 2025 – Sep. 2025*

- Engineered a Transformer-based sequential recommendation engine for all-modality generative retrieval, effectively fusing massive sparse ID features with dense multi-modal embeddings.
- Optimized sequence representation learning utilizing an InfoNCE contrastive loss framework with dynamic negative sampling, improving model discrimination on highly sparse user behaviors.
- Accelerated the end-to-end inference pipeline by integrating **FlashAttention** for memory-efficient sequence modeling and implementing a high-throughput batched Approximate Nearest Neighbor (ANN) search via PyTorch.

High-Performance Sokoban AI Solver Course Project
C++ Algorithm Developer *Nov. 2025 – Dec. 2025*

- Engineered a high-performance puzzle solver using **A* search**, optimizing memory footprint via bit-level state representation and a custom hash function for ultra-fast state deduplication.
- Mitigated state space explosion and search latency by implementing aggressive pruning strategies, including reverse-reachability static deadlock precomputation and dynamic deadlock detection.

Technical Skills

Languages: C/C++, Python
Machine Learning: PyTorch, Transformers, LLM Inference, Speculative Decoding
Systems & Architecture: ML Compilers, High-Performance Computing (HPC), Memory Optimization
Developer Tools: Linux, Git, GDB, LaTeX