JINGWEI ZUO

Rice University, Houston, TX
+1 (281) 827-4384 | e: Jingwei.Zuo@rice.edu | https://jingwei-zuo.com/

EDUCATION

Rice University — Houston, TX, USA
Doctoral Study in Computer Science Department — Aug. 2025-Now
- Working with Prof. Yuke Wang, focusing on Machine Learning Systems

Tsinghua University — Beijing, China
B.Sc. in Mathematics and Physics & B.Eng. in Electrical Engineering — Sept. 2021-June 2025
- Working with Prof. Zhiyuan Liu

RESEARCH INTERESTS

- Efficient inference and training/fine-tuning system of Large Language Models
- Algorithm-system codesign of modern generative AI models

PUBLICATIONS

DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads
Guangxuan Xiao, Jiaming Tang, **Jingwei Zuo**, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, Song Han.
arXiv:2410.10819, NeurIPS 2025 Poster

RESEARCH AND WORK EXPERIENCES

Moonshot AI — Beijing, China
AI Infrastructure Intern — Dec. 2024 – May 2025
- Working on the inference optimization of Large Language Models
- Benchmarking, profiling the inference process by Nsight System or PyTorch profiler to seek space for optimization methods such as speculative decoding, model parallelism, and effective GPU scheduling
- Providing inference support to the pretrained models and actively seek better algorithms, like MoE or speculative decoding

CMU Infinite Lab — Remotely
Research Assistant to Prof. Beidi Chen — June-Oct. 2024
- Conducted research on accelerating long-context language model (LLM) inference, using top-k sparse attention to support long context generation with minimal latency
- Explore dapproximate nearest neighbor search (ANNS) to retrieve the key-value pairs with the largest attention score, thereby reducing GPU memory usage

MIT Han Lab — Cambridge, MA, USA
Research Assistant to Prof. Song Han — Oct. 2023-May 2024
DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads
- Designed a novel framework that significantly reduces computational memory and latency in long-context large language models
- Devise a method that applies full Key-Value (KV) caching to Retrieval Heads while employing a constant-length KV cache for other heads (*Streaming Heads*)
- Realize up to 2.12× reduction in inference memory and up to 3.05× acceleration in decoding for models like Llama-2/3 and Mistral, with minimal accuracy loss

Tsinghua Natural Language Processing Lab — Beijing, China
Research Assistant to Prof. Zhiyuan Liu — Mar. 2023-Aug. 2023
AGENTVERSE: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors
- Co-designed a cutting-edge AI framework enabling *multiple agents* to *collaborate* like human teams
- Validated the framework's effectiveness in diversified circumstances such as reasoning, coding, tool utilization, and embodied AI.
- Built and released the code on GitHub, with a gain of *4.8K stars* so far.

PROJECTS

1. **NeRF Octree Optimization** — June 2023
- Utilized *Octree* data structure to optimize the memory consumption and time efficiency of NeRF rendering
- Attained up to *4x* memory optimization compared to *voxel* storage and the rendering time is equivalent
- Utilized PyTorch and the obtained the basic idea to make an AI model more efficient

2. **Wordinary: Comprehensive Learning Suite for Language Learners** — July 2021-Feb. 2022

- Created a multifaceted educational software designed to enhance *vocabulary building* for English learners, focusing on *high-frequency word extraction, quiz generation,* and *standard pronunciation audio creation*
- Engineered the software using Python 3 for backend processing and C# .NET for a user-friendly interface, ensuring compatibility with Windows systems
- Actively managed and updated the project on GitHub, demonstrating continuous improvement and engagement with the user community

SELECTED AWARDS AND HONORS

- **Dean's List** 2023Fall
  Issued by College of Engineering, Northeastern University
- **Academic Excellence Scholarship** 2022-2023
  Issued by Tsinghua University
- **Comprehensive Scholarship** 2021-2022
  Issued by Tsinghua University

SKILLS

- Proficient in Python with strong experience of using NumPy, MatPlotLib, PyTorch, HuggingFace, NSight-Systems.
- Advanced coding skills, proficient in C, C++, and Python
- Professional English (TOEFL: 110, R30 L30 S26 W24) and native in Chinese