

JINGWEI ZUO

Tsinghua University, P.R. China
+86 159-5290-6186 | e: naohzjw@gmail.com

EDUCATION

Tsinghua University

B.Sc. in Mathematics and Physics & B.Eng. in Electrical Engineering (dual degree)

- GPA: 3.88/4.00
- Awarded Scholarship for Academic Excellence 2022 – 2023

Beijing, China
Sept. 2021 – June 2025

Northeastern University

Exchange Student at College of Engineering

- GPA: 4.00
- Selected on Dean's List

Boston, MA, USA
Sept. – Dec. 2023

PUBLICATIONS & PREPRINT

AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors

Weize Chen, Yusheng Su, **Jingwei Zuo**, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, Zhiyuan Liu, Maosong Sun, Jie Zhou. *In Proceedings of ICLR, 2024*

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. *Under review of ICLR, 2025*

RESEARCH EXPERIENCE

Carnegie Mellon University (Infinite Lab)

Research Assistant to Prof. Beidi Chen

- Conducted research on accelerating long-context language model (LLM) inference, targeting efficient attention mechanisms to support extended context windows with minimal latency
- Explored approximate nearest neighbor search (ANNS) to retrieve the key-value pairs with the largest attention score thereby reducing GPU memory usage
- Conducted experiments to compare the latency and recall rate of different ANNS methods and their performance dealing with token embeddings
- Implemented the end-to-end pipeline and tested our method's performance on benchmarks like GSM8K and RULER

Remote Work
June – Oct. 2024

Massachusetts Institute of Technology (Han Lab)

Research Assistant to Prof. Song Han

DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads

- Pioneered a novel framework that significantly reduces computational memory and latency in long-context large language models
- Engineered a lightweight, optimization-based algorithm utilizing synthetic data to accurately identify the *Retrieval Heads*
- Devised a method that applies full Key-Value (KV) caching to Retrieval Heads while employing a constant-length KV cache for other heads (*Streaming Heads*)
- Realized up to $2.12\times$ reduction in inference memory and up to $3.05\times$ acceleration in decoding for models like Llama-2/3 and Mistral, with minimal accuracy loss

Cambridge, MA, USA
Oct. 2023 – May 2024

Tsinghua University (THU Natural Language Processing Lab)

Research Assistant to Prof. Zhiyuan Liu

AGENTVERSE: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors

- Co-designed a cutting-edge AI framework enabling *multiple agents* to *collaborate* like human teams
- Designed the *dynamic role assignment* strategy
- Validated the framework's effectiveness in diversified circumstances such as reasoning, coding, tool-utilization, and embodied AI, etc.
- Revealed *emergent sociological behaviors* such as volunteer behaviors and conformity behaviors
- Built and released the code at <https://github.com/OpenBMB/AgentVerse>

Beijing, China
March 2023 – Aug. 2023

PROJECT EXPERIENCES

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 was equivalent

- Utilized PyTorch and the obtained the basic idea to make an AI model more efficient

Markov Chain Application in Tennis Competitions

Dec. 2022

- Course project of *Probability and Stochastic Processes*, merged my passion for tennis with mathematical analysis.
- Utilized *Markov Chain* analysis to demonstrate the *stabilizing effect* of tennis's multi-game per set and multi-point per game rules on player performance.

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
- Innovated by introducing customizable features for varied educational needs, such as setting benchmarks for word extraction adaptable for exams like CET-4, TOEFL, or GRE
- Actively managed and updated the project on [GitHub](#), demonstrating continuous improvement and engagement with the user community

SELECTED AWARDS AND HONORS

- **Comprehensive Scholarship** 2021 – 2022
Issued by Tsinghua University
- “TI Cup” Digital System Innovation Design Competition (Third Prize) Oct. 2022
Designed self-tracking algorithms on microcontrollers and also intelligent algorithms to find the best route
- “Xindong” Vehicle Competition (Third Prize) Jan. 2022
Developed a self-tracking mini-vehicle using a microcontroller, incorporating PID control methods and camera-based tracking for enhanced autonomous navigation
- Earn outstanding award in *Software Programming Training*, *Android Programming*, and *Embedded System Design* courses
- National Olympiad in Informatics in Provinces (Second Prize) Dec. 2018

SKILLS

- Proficient in Python with three years experience of using numpy, matplotlib, and pytorch
- Advanced coding skills, proficient in developing complex algorithms and solutions across multiple programming languages such as C, C++, C#, Java, and Python
- Professional fluency in English (TOEFL: 110, R30 L30 S26 W24) and native Chinese speaker
- Three years experience playing tennis