

# ZHIIHAO ZHANG

+1 949-981-1208  $\diamond$  zzh\_jackfram@outlook.com

## EDUCATION

---

**Carnegie Mellon University, USA**

*2022.9 - Now*

Ph.D. in Computer Science

**Carnegie Mellon University, USA**

*2020.9 - 2022.8*

Master of Science in Robotics (MSR)

**Renmin University, China**

*2016.9 - 2020.7*

B.S. in Computer Science

## RESEARCH EXPERIENCES

---

**Carnegie Mellon University, Catalyst**

*Pittsburgh, U.S*

Research Assistant, advised by Prof. Zhihao Jia

*2021.3-now*

- Machine Learning System

**Carnegie Mellon University, Intelligent Control Lab**

*Pittsburgh, U.S*

Research Assistant, advised by Prof. Changliu Liu

*2020.9-2021.3*

- Deep learning theory related topics, eg. Neural Tangent Kernel, Rademacher Complexity, Norm Based NN Capacity Measurement.

**University of California Berkeley, Mechanical Systems Control Lab**

*Berkeley, U.S*

Research Intern, advised by Prof. Masayoshi Tomizuka

*2019.10-2020.3*

- “Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network”, an interactive trajectory prediction method using GNN framework

## PUBLICATIONS

---

- **TidalDecode: Fast and Accurate LLM Decoding with Position Persistent Sparse Attention**, under review.  
Lijie Yang\*, **Zhihao Zhang\***, Zhuofu Chen, Zikun Li, Zhihao Jia
- **Accelerating Retrieval-augmented Language Model Serving with Speculation**, International Conference on Machine Learning (ICML), July 2024.  
**Zhihao Zhang**, Alan Zhu, Lijie Yang, Bruce Xu, Lanting Li, Phitchaya Mangpo Phothilimthana, Zhihao Jia
- **Communication Bounds for the Distributed Expert Problem**, The Conference on Neural Information Processing Systems (NeurIPS), December 2024.  
Zhihao Jia, Qi Pang, Trung Tran, David Woodruff, **Zhihao Zhang**, Wenting Zheng (alphabetic order)
- **SpecInfer: Accelerating Generative LLM Serving with Speculative Inference and Token Tree Verification**, In Proceedings of the ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), April 2024.  
Xupeng Miao\*, Gabriele Oliaro\*, **Zhihao Zhang\***, Xinhao Cheng\*, Zeyu Wang, Rae Ying Yee Wong, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, Zhihao Jia
- **GradSign: Model Performance Inference with Theoretical Insights**, The Tenth International Conference on Learning Representations (ICLR 2022)  
**Zhihao Zhang**, Zhihao Jia
- **Social-WaGDAT: Interaction-aware Trajectory Prediction via Wasserstein Graph Double-Attention Network**, IEEE Transactions on Intelligent Transportation Systems (TITS)  
Jiachen Li, Hengbo Ma, **Zhihao Zhang**, Masayoshi Tomizuka

## AWARDS

---

- Meta Research Award, 2022
- Google Faculty Research Award, 2022