

Zining Song

M.S. Student, ShanghaiTech University × MetaX Joint Program

✉ zining.song@outlook.com 🏠 bio.zining-song.com 🎓 Google Scholar 🐼 laowoniu-snail

Research Interests

Deep learning and GPU performance engineering. I am keen to explore new models, architectures, and algorithms, and to optimize existing algorithms by aligning them with the compute characteristics of GPU hardware.

Education

ShanghaiTech University — M.S. (Professional Master’s), Computer Science & Technology Sep 2025 – Present
Joint training program with MetaX Integrated Circuits

- Selected coursework: Advanced Distributed Systems, Network & Information Security, Natural Language Processing, Algorithm Design & Analysis.

ShanghaiTech University — B.S., Computer Science & Technology Sep 2021 – Jun 2025

- Coursework: Algorithms & Data Structures, Introduction to Machine Learning, Matrix Computation, Parallel Computing.

Experience

MetaX Integrated Circuits — LLM Algorithm Intern Jun 2025 – Sep 2025

- Surveyed methods for large-model quantization, deployment, evaluation, and fine-tuning.
- Quantized the natively-FP8 **DeepSeek-R1-0528** to INT8 (W8A8) with open-source tooling to exploit the INT8 hardware compute acceleration of the MetaX C500 GPU.
- Deployed the INT8 DeepSeek-R1-0528 on a 2-node, 8-GPU MetaX C500 server and validated end-to-end serving.
- Researched model-evaluation methodology to quantify the quality gap of the quantized model versus the original across multiple dimensions.
- Fine-tuned the LLM with LoRA to improve the performance of the quantized model, and produced related case studies and tutorials.
- Authored a Gitee AI × MetaX large-model API orchestration example and tutorial.

Publications

FreeCap: Hybrid Calibration-Free Motion Capture in Open Environments AAAI 2025

Aoru Xue*, Yiming Ren*, **Zining Song**, Mao Ye, Xinge Zhu, Yuexin Ma. [arXiv:2411.04469](#)

FreeCap is a calibration-free system that uses LiDAR and movable cameras to capture accurate multi-person motion in open environments, and scales to additional cameras for improved accuracy. My primary contribution was assisting in data collection and leading method evaluation: I designed and conducted a series of comparative and ablation experiments that validated the effectiveness and superiority of the proposed method.

HOI-M³: Capture Multiple Humans and Objects Interaction within Contextual Environment CVPR 2024 (Highlight)

Juze Zhang*, Jingyan Zhang*, **Zining Song**, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, Jingya Wang. [arXiv:2404.00299](#)

HOI-M³ contributes a high-quality, large-scale multi-person and multi-object interaction dataset for motion capture and downstream research such as motion generation and embodied intelligence. My primary contribution was modeling and tracking object positions and poses, adopting a method that fuses IMU and multi-camera data and uses IMU measurements as a prior for object rotation.

Skills

Programming	C/C++, Python; familiar with GPU/CUDA parallel programming, MATLAB, and $\LaTeX$.
LLM & Systems	Model quantization (FP8/INT8, W8A8), large-model deployment & serving, evaluation, and fine-tuning.
Deep Learning	Transformer, ResNet, PointNet, GAN/CycleGAN; GPU performance optimization.