Runyu Lu

runyulu@umich.edu ◇ LRY89757 ◇ runyulu.com ◇ runyu-lu ◇ +1 (734) 216-0351

EDUCATION

PhD, Computer Science Sept. 2024 - Present
University of Michigan, Research Area: Machine Learning System. Advisors: Mosharaf Chowdhury and Ang Chen

Bachelor Engineering, Computer Science Sept. 2020 - June. 2024
Huazhong University of Science and Technology, Computer Excellence Program, GPA: 3.95 / 4.00

PUBLICATIONS

DSA: Efficient Inference For Video Generation Models via Distributed Sparse Attention **ICLR 2026**
Shenggui Li, **Runyu Lu**, Qiaoling chen, Haiyan Yin, Yueming Lyu, Yonggang Wen, Ivor Tsang, Tianwei Zhang
Keywords: DiT Inference, Parallelization, Sparse Attention

TetriServe: Efficiently Serving Mixed DiT Workloads **ASPLOS 2026**
Runyu Lu*, Shiqi He*, Wenxuan Tan, Shenggui Li, Ruofan Wu, Jeff J. Ma, Ang Chen, Mosharaf Chowdhury
Keywords: DiT Serving, GPU Resource Scheduling

MuxServe: Flexible Multiplexing for Efficient Multiple LLM Serving **ICML 2024**
Jiangfei Duan, **Runyu Lu**, Haojie Duanmu, Xiuhong Li, Xingcheng ZHANG, Dahua Lin, Ion Stoica, Hao Zhang
Keywords: LLM Serving, GPU SM Sharing

White-box Compiler Fuzzing Empowered by Large Language Models **OOPSLA 2024**
Chenyuan Yang, Yinlin Deng, **Runyu Lu**, Jiayi Yao, Jiawei Liu, Reyhaneh Jabbarvand, Lingming Zhang
Keywords: LLM for Compiler, LLVM, LLVM IR Fuzzing

ARXIV

ModalGlue: Distributed Multimodal Training Must Be Multimodality-Aware
Insu Jang, **Runyu Lu**, Nikhil Bansal, Ang Chen, Mosharaf Chowdhury
Keywords: MultiModal Training, FlexAttention

WORK EXPERIENCE

Isaac GR00T N*: RL Post Training on GR00T Foundation Model.* June.2025 - Present, **NVIDIA GEAR**

ColossalAI*: Design and Implement Diffusion Inference Engine* March.2024 - July.2024, **HPC-AI**

LLVM*: Optimize GPU Compiler Backend* April 2023 - Sept.2023, **SenseTime**

NCNN*: Develop High Performance CNN Inference Engine* April 2022 - Nov.2022, **Tencent**

SKILLS

Languages: Native in Mandarin; Fluent in English

Last Update 2026.03.16