

LESHU LI (李乐书)

 [Personal-Website](#)  [E-mail](#)  [Linkedin](#)  [Github.com](#)  [Google Scholar](#)

Education

New York University, Courant

Ph.D. in Computer Science

Sep. 2026 – Sep. 2031 (Expected)

New York, NY, USA

University of Minnesota, Twin Cities

M.S. in Electrical and Computer Engineering

Sep. 2023 – May 2026

Minneapolis, MN, USA

Sichuan University

B.Eng. in Telecommunications Engineering

Sep. 2019 – Jul. 2023

Chengdu, Sichuan, China

Research Areas & Focus

My research interests lie at the intersection of Machine Learning Systems and Computer Architecture, with a strong emphasis on **Efficient AI** and **Hardware-Software Co-design**. Specifically, I focus on efficient large language model (LLM) inference, particularly Speculative Decoding. I also explore efficient video understanding and streaming inference with V-JEPA, acceleration of 3D Gaussian Splatting (3DGS), and performance and energy optimization for Augmented and Virtual Reality (AR/VR) systems. My broader goal is to optimize AI systems across algorithms, system software, and hardware architectures.

Publications & Manuscripts

1. **L. Li***, J. Qin*, P. Jie, Z. Wan, H. Qu, Y. Han, P. Zhen, H. Zhang, Y. Cao, T. Cheng, Y. Zhao. **RTGS: Real-Time 3D Gaussian Splatting SLAM via Multi-Level Redundancy Reduction**. *58th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2025. *Equal contribution.
2. **L. Li**, P. Jie, Y. Zhao. **Pocket-SLAM: Rendering-Area-Aware Pruning for Memory-Efficient 3DGS-SLAM**. *International Conference on Robotics and Automation (ICRA)*, 2026.
3. **L. Li**, A. Liu, C. Zhao, S. Zhang. **LipoAgent: Coordinating Fine-Tuned LLM Agents for Safer Lipid Design**. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.
4. Z. Ye, Y. Fu, J. Zhang, **L. Li**, Y. Zhang, S. Li, C. Wan, C. Wan, C. Li, S. Prathipati, Y. Lin. **Gaussian Blending Unit: An Edge GPU Plug-in for Real-Time Gaussian-Based Rendering in AR/VR**. *31th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.
5. Z. Ye, C. Wan, C. Li, J. Hong, S. Li, **L. Li**, Y. Zhang, Y. C. Lin. **3D Gaussian Rendering Can Be Sparser: Efficient Rendering via Learned Fragment Pruning**. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
6. Z. Ye, Z. Wang, K. Xia, J. Hong, **L. Li**, L. Whalen, C. Wan, Y. Fu, Y. C. Lin, S. Kundu. **LAMB: A Training-Free Method to Enhance the Long-Context Understanding of SSMS via Attention-Guided Token Filtering**. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.
7. **L. Li**, F. Yu, B. McDanel, S. Q. Zhang. **DSD: A Distributed Speculative Decoding Solution for Edge-Cloud Agile Large Model Serving**. *arXiv preprint*, 2026.
8. H. Wang, Y. Wang, **L. Li**, Y. Ren, S. Q. Zhang. **LASER: Loss-Aware Singular-value Decomposition and Rank Allocation for Efficient Low-Precision Vision-Language Models**. *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Industry Internship Experience

REDstar@hi Lab, Xiaohongshu Technology Co., Ltd (REDnote)

LLM Pre-training Intern

Sep. 2025 – Feb. 2026

Shanghai, CN

- Led systematic benchmarking of advanced attention mechanisms including **MLA**, **MFA**, and **GQA**, analyzing trade-offs in memory footprint, training throughput, and convergence stability under large-scale pre-training settings.

- Contributed to the design and partial implementation of **mHC (Manifold-Constrained Hyper-Connections)**, introducing manifold-regularized high-order connectivity to enhance representational capacity while improving parameter efficiency and training stability.
- Optimized large-scale training on NVIDIA Blackwell GPU clusters, improving achieved TFLOPs utilization through kernel tuning, parallelism reconfiguration (Megatron/FSDP), and communication-computation overlap optimization.
- Collaborated with infrastructure engineers to diagnose and mitigate system bottlenecks across data pipelines, model parallelism strategies, and distributed communication stacks during training and inference.

Research Experience

Sai Lab, New York University

Mar. 2026 – Present

PhD Researcher | Advisors: Prof. Sai Qian Zhang

New York, NY, USA

- **Efficient LLM Inference via Tree-Based Speculative Decoding**

- * Developed a tree-based speculative decoding framework to accelerate large language model (LLM) inference through parallel token verification and efficient draft generation.
- * Designed acceptance-length-aware training objectives and optimized draft token selection and tree construction to improve expected accepted length and inference efficiency.
- * Evaluated speculative decoding strategies across Qwen and Llama model families, analyzing acceptance length, decoding throughput, and inference latency under different draft-target configurations.

- **Efficient Streaming Video Understanding with V-JEPA**

- * Developed an efficient streaming video understanding framework based on V-JEPA for real-time action prediction using visual inputs and multimodal signals.
- * Designed a streaming KV-cache mechanism with attention-based token pruning and selective retention to reduce redundant computation and improve inference efficiency.
- * Optimized temporal feature processing and lightweight prediction heads to balance action prediction accuracy, computational overhead, and end-to-end latency.

Zhao Lab, University of Minnesota, Twin Cities

Jun. 2024 – May 2025

Research Assistant | Advisor: Prof. Yang (Kaite) Zhao

Minneapolis, MN, USA

- **Efficient 3D Gaussian Splatting (3DGS) SLAM on Edge Devices**

- * Profiled rendering and training bottlenecks in 3DGS-based SLAM, identifying computational imbalances, memory access conflicts, and inefficient workload distribution on resource-constrained devices.
- * Optimized hardware architecture and workload scheduling to improve rendering throughput, computational efficiency, and real-time SLAM performance on edge GPUs.
- * Developed a Gaussian pruning strategy with dynamic memory budgeting to reduce peak GPU memory consumption during training and inference while preserving rendering quality and SLAM accuracy.

EIC Lab, Georgia Institute of Technology

Mar. 2024 – May 2025

Research Intern | Advisor: Prof. Yingyan (Celine) Lin

Atlanta, GA, USA

- **Efficient Inference for Diffusion Large Language Models (dLLMs)**

- * Developed a selective remasking algorithm to identify and correct incorrectly decoded tokens during iterative generation, improving correctness by 23% on HumanEval and GSM8K benchmarks.
- * Proposed a confidence-trend-based early decoding strategy and investigated memory-efficient mechanisms to reduce redundant token updates and computational overhead in diffusion LLM inference.

Technical Skills

Programming Languages: CUDA, Triton, C, C++, Python, Verilog

EDA Tools: Virtuoso, Innovus, Vivado, HSPICE

Honors & Awards

- **Scholarship:** The Comprehensive Second Class Scholarship in Sichuan University(2019-2020)
- **Grant:** MICRO 2025 Travel Grant (\$580)