LESHU LI（李乐书）

☎ +1 6519004128 ✉ lileshu0412@gmail.com in www.linkedin/leshu github.com/Nemo0412

Education

SICHUAN University **Sep. 2019 – July. 2023**
Bachelor of Telecommunications Engineering *Chengdu, Sichuan, China*

University of Minnesota, Twin Cities **Sep. 2023 – Mar. 2026 (Expected)**
Master of Electrical and Computer Engineering ***(GPA: 3.9/4.0)*** *Minneapolis, MN, USA*

Research Areas & Focus

My research interests lie at the intersection of computer architecture and machine learning systems, with a strong emphasis on ***Hardware–Software Co-design*** for ***Efficient AI***. Specifically, I focus on optimizing 3D Gaussian Splatting (3DGS) and inference for large language models (LLMs). I am also actively studying the pre-training stage of LLMs and am highly interested in developing optimization techniques for this phase. Overall, I am passionate about advancing efficient AI across algorithms, architectures, and hardware.

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. 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.
3. 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.
4. 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.
5. **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 (submitted).

Research Experience

Zhao Lab, University of Minnesota, Twin Cities **Jun. 2024 – Present**
Research Assistant | Advisors: Prof. Yang (Kaite) Zhao *Minneapolis, MN, USA*

- **Conducting Research on Optimizing 3D Gaussian Splatting (3DGS)-Based SLAM for Edge Devices to Achieve Real-Time Performance**
  - Analyzed and profiled bottlenecks in current 3DGS-based SLAM systems, focusing on factors causing run-time inefficiencies and computational limits.
  - Designed and implemented solutions to address computational imbalances in rendering and training, resolving memory conflicts, and workload distribution issues.
  - Evaluated the optimized hardware architecture on a edge device, showing significant improvements in speed, efficiency, and resource usage compared to prior implementations.
- **Developing Memory-Efficient 3DGS SLAM for Large-Scale Outdoor Deployment**
  - Investigated memory consumption of 3DGS SLAM from both static storage (keyframe management) and dynamic usage (peak GPU memory), aiming to enable real-time operation on edge devices.
  - Proposed a rendering-area–based pruning strategy to reduce peak memory usage during training and inference, significantly improving scalability and efficiency.
  - Designed a tile-level budgeting mechanism, where the budget is computed based on Gaussian gradients, to ensure that pruning reduces memory usage without degrading the final SLAM accuracy.

EIC Lab, Georgia Institute of Technology **Mar. 2024 – May. 2025**
Research Intern | Advisor: Prof. Yingyan (Celine) Lin *Atlanta, GA, USA*

- **Improving 3D Gaussian Splatting (3DGS) Rendering Performance on Edge Devices**
  - ∗ Profiled the 3DGS rendering algorithm using Nvidia Nsight Systems on the Orin NX GPU and identified significant redundant computations in the current approach.
  - ∗ Co-designed a dedicated hardware module that can seamlessly integrate into existing GPU architectures to improve data locality and leverage Gaussian reuse caches for optimized rendering.
  - ∗ Implemented the proposed Gaussian Blending Unit (GBU) in Verilog, synthesized the design into a gate-level netlist with Cadence Genus in TSMC 28nm, and built a cycle-accurate simulator on top of GPGPU-Sim to evaluate rendering throughput when integrated with edge devices.
- **Enhancing Diffusion Large Language Models (dLLM & Llada)**
  - ∗ Designed a novel *remask* algorithm that selectively overwrites incorrectly decoded tokens during dLLM decoding, improving correctness by 23% on *HumanEval* and *GSM8K*.
  - ∗ Proposed a confidence-trend-based early decoding strategy to accelerate the remask process, enabling efficiency gains without sacrificing decoding latency or accuracy.
  - ∗ Investigated memory-efficient mechanisms to reduce redundancy during iterative decoding, improving scalability of diffusion-based LLMs on constrained hardware.

Sai Lab, New York University **Aug. 2025 – Present**
Research Intern | Advisors: Prof. Sai Qian Zhang & Prof. Bradley McDanel *New York, NY, USA*

- **Designing a Speculative Distributed LLM System for Efficient Inference**
  - ∗ Participated in the design of a distributed LLM inference framework that adapts to both the diversity of incoming prompts and the heterogeneity of underlying devices.
  - ∗ Developed a dynamic scheduling strategy that assigns requests based on prompt complexity and device capabilities, improving overall throughput and resource utilization.
  - ∗ Integrated speculative decoding into the distributed system to accelerate inference, leveraging early-exit predictions to reduce latency while maintaining accuracy.
- **Efficient Generative AI and Human-in-the-Loop Strategies in Drug Delivery**
  - ∗ Fine-tuned the Qwen3-235B large language model to develop a generative AI algorithm for optimizing drug delivery design, improving both prediction accuracy and computational efficiency.
  - ∗ Incorporated a human-in-the-loop framework that integrates domain knowledge, including evaluating drug synthesizability and identifying low-confidence predictions from the LLM, to guide model decisions.
  - ∗ Validated the proposed approach through simulations and case studies, and further demonstrated its effectiveness via laboratory drug synthesis and cell experiments, confirming the model's ability to accurately predict drug efficiency and showcasing its potential for real biomedical applications.

Internship Experience

REDstar@hi Lab, Xiaohongshu Technology Co., Ltd (REDnote) **Sep. 2025 – Present**
Internship in LLM Pre-training *Shanghai, CN*

- Collaborated with the Infra team to address training and inference bottlenecks, contributing to the architectural design and iterative optimization of large language models (LLMs).
- Explored efficient scaling strategies such as context scaling and parameter scaling, and investigated next-generation GPU-friendly model architectures and algorithms.
- Investigated advanced mechanisms including Attention, MoE, optimizer strategies, and novel learning paradigms using Megatron and FSDP frameworks.

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)