李翔宇

电话/微信: (+86)15910912361 | Email: xiangyu.sdlc@foxmail.com
个人主页：https://xxxxyu.github.io | Google Scholar: Xiangyu Li | GitHub: xxxxyu

教育经历

清华大学 | 智能产业研究院（AIR） | *博士研究生* 2022.09—2027.06

研究方向：端侧智能，具身智能，机器学习系统。导师：刘云新教授。预计 2027 年毕业。

曾获清华之友-智能产业研究院清智奖学金（2025.12），清华之友-济宁英才奖学金（2024.11）。

清华大学 | 电子工程系 | *工学学士* 2018.09—2022.06

研究方向：图挖掘系统。曾任电子系学生科协软件部部长、副主席。

曾获清华大学社会工作优秀奖学金（2021.12），第三十六届全国部分地区大学生物理竞赛特等奖（2019.12）。

研究经历

ActProbe / EmbodiSkill: 具身智能体失败检测与反思优化 | *共同一作 / 参与* 2026.02—至今

- 研究机器人策略早期失败检测与技能级反思闭环；ActProbe：论文/代码/主页；EmbodiSkill：论文。

OxyGen: 面向多任务并行 VLA 推理的统一 KV Cache 管理 | *第一作者* 2025.07—2026.03

- 研究动机：具身智能体需要具备多任务并行推理的能力，但已有方法对内存与计算资源需求高，难以在端侧部署。
- 主要贡献：提出统一的 KV Cache 管理方法，实现跨任务复用与跨帧持续解码，并基于 openpi 框架完成系统实现。
- 主要结果：在单卡 4090 GPU 上实现至多 **3.7**× 加速，可同时达到 **200 tokens/s** 文本吞吐与 **70 Hz** 动作频率。
- 论文：https://arxiv.org/abs/2603.14371 ；代码：https://github.com/air-embodied-brain/OxyGen 。

Vec-LUT: 基于向量查找表的低比特 LLM 推理加速 | *共同一作*（**MobiSys 2026**） 2024.07—2025.05

- 研究动机：大模型端侧部署中低比特量化是刚需，但硬件缺乏混合精度支持，实际加速与理论上限有显著差距。
- 主要贡献：提出基于向量查找表（Vector Lookup Table）的混合精度矩阵乘范式，并基于 llama.cpp 框架实现系统。
- 主要结果：对于三值 LLM，在移动设备与边缘服务器上实现至多 **4.2**× 端到端推理加速；CPU 双核性能超过 NPU。
- 论文：https://dl.acm.org/doi/10.1145/3745756.3809200 ；代码：https://github.com/OpenBitSys/vlut.cpp 。

FlexNN: 面向内存受限的自适应模型内存管理 | *第一作者*（**MobiCom 2024**） 2022.01—2023.08

- 研究动机：内存限制是端侧模型部署的主要瓶颈（即"能否放得下"），但模型不能无限压缩，需要系统层优化。
- 主要贡献：提出张量级的细粒度层切分-加载-计算联合规划方法，并基于腾讯 ncnn 框架完成系统实现。
- 主要结果：在手机与嵌入式设备上实现至多 **93.8%** 内存节省，同时仅增加 **3.6%** 推理延迟；适应开销小于 1 秒。
- 论文：https://dl.acm.org/doi/10.1145/3636534.3649391 ；代码：https://github.com/xxxxyu/FlexNN 。

工作经历

字节跳动 | 产品研发和工程架构部 | *大数据架构实习生* 2021.06—2021.09

- **参与 FaaS（Function-as-a-Service）基础设施研发**：支持字节上游业务的各类高并发与弹性扩缩容负载。
- **独立完成节点镜像更新与缓存策略优化**：将 worst case 下冷启动开销降低 **3∼4 个数量级**，显著提升线上系统 QoS。

项目经历

海信合作项目：云侧与端侧 LLM/VLM 推理加速 | *负责项目主要工作* 2025.09—2026.06

- **负责垂域 LLM 的低比特训练与推理算子优化**：准确率与推理速度均达业界 SOTA（已完成）。INT2 量化下相比 FP16 准确率损失小于 **1%**，在 A40 GPU 上相比 INT4 至多加速 **1.9**×。将支持海信线上业务并开源（进行中）。
- **负责端侧 VLM 的多模态压缩**：探索视觉 token 压缩方法并在端侧 NPU 上部署（进行中）。

宝马合作项目：面向智能座舱的轻量 LLM 定制与部署 | *负责端侧部署部分* 2023.12—2024.04

- **负责高通芯片上的 LLM 部署测试**：通过部署量化小模型，达到 **5.2**× 内存节省，**6.4**× 推理加速。

技能与工具

- 熟悉 **C/C++** 和 **Python** 编程语言；熟悉各类端侧设备与 Linux 环境。
- 熟悉主流推理部署框架，有丰富的基于 **vLLM**，**llama.cpp**，**ncnn** 等框架的开发与部署经验；有开源贡献经历。
- 了解 PyTorch 生态下的常用训练框架与工作流，有一定的模型微调与蒸馏经验；乐于接触更大规模训练。

论文列表

[1] **Xiangyu Li**. 2026. Building Efficient Inference Systems for Resource-Constrained Edge AI Deployment. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services Companion (MobiSys Companion 2026).*
MobiSys 2026 Rising Stars Forum. *Paper: ACM DL.*

[2] Bingjia Huang*, **Xiangyu Li***, Xiang Wang, Liang Mi, Zixu Hao, Weijun Wang, Hao Wu, Kun Li, Yunxin Liu, and Ting Cao. 2026. ActProbe: Action-Space Probe for Early Failure Detection of Generative Robot Policies. *arXiv preprint arXiv:2606.08508.*
Paper: arXiv. Code: GitHub. Page: Website.

[3] Ruofei Ju*, Xinrui Wang*, Xin Ding, Yifan Yang, Hao Wu, Shiqi Jiang, Qianxi Zhang, Hao Wen, **Xiangyu Li**, Weijun Wang, Kun Li, Yunxin Liu, Haipeng Dai, Wei Wang, and Ting Cao. 2026. EmbodiSkill: Skill-Aware Reflection for Self-Evolving Embodied Agents. *arXiv preprint arXiv:2605.10332.*
Paper: arXiv.

[4] **Xiangyu Li**, Huaizhi Tang, Xin Ding, Weijun Wang, Ting Cao, and Yunxin Liu. 2026. OxyGen: Unified KV Cache Management for VLA Inference under Multi-Task Parallelism. *arXiv preprint arXiv:2603.14371.*
Paper: arXiv. Code: GitHub.

[5] **Xiangyu Li***, Chengyu Yin*, Weijun Wang, Jianyu Wei, Ting Cao, and Yunxin Liu. 2026. Vec-LUT: Vector Table Lookup for Parallel Ultra-Low-Bit LLM Inference on Edge Devices. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys 2026).*
MobiSys 2026 (CCF-B, THCPL-A). Featured Paper *for On-Device AI session.*
Paper: ACM DL. Code: GitHub. Model: Hugging Face.

[6] Yi Sun, Han Wang, Jiaqiang Li, Jiacheng Liu, **Xiangyu Li**, Hao Wen, Yizhen Yuan, Huiwen Zheng, Yan Liang, Yuanchun Li, and Yunxin Liu. 2025. An Empirical Study of LLM Reasoning Ability Under Strict Output Length Constraint. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025).*
EMNLP 2025 (CCF-B, THCPL-A) main conference. *Paper: arXiv. Page: Website.*

[7] Xiang Wang, Lingxiao Ma, Ziyan Fu, **Xiangyu Li**, Yuanchun Li, Ju Ren, Yaoxue Zhang, and Yunxin Liu. 2025. Squeezer: Efficient Multi-DNN Inference for Edge Video Analytics via Cross-Model Scheduling. *IEEE Transactions on Mobile Computing (TMC).*
TMC (CCF-A, THCPL-A). *Paper: IEEE Xplore.*

[8] Jiacheng Liu, Yuanchun Li, Liangyan Li, Yi Sun, Hao Wen, **Xiangyu Li**, Yao Guo, and Yunxin Liu. 2024. ChainStream: An LLM-based Framework for Unified Synthetic Sensing. *arXiv preprint arXiv:2412.15240.*
Paper: arXiv. Code: GitHub. Page: Website.

[9] **Xiangyu Li**, Yuanchun Li, Yuanzhe Li, Ting Cao, and Yunxin Liu. 2024. FlexNN: Efficient and Adaptive DNN Inference on Memory-Constrained Edge Devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom 2024).*
MobiCom 2024 (CCF-A, THCPL-A). *Paper: ACM DL. Code: GitHub. Slides: Link.*

[10] Yuanchun Li†, Hao Wen‡, Weijun Wang‡, **Xiangyu Li**‡, Yizhen Yuan‡, Guohong Liu‡, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanjing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. 2024. Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security. *arXiv preprint arXiv:2401.05459.*
Survey & Position (400+ citations). *"Efficiency" section lead. Paper: arXiv. Code: GitHub.*

[11] Guohao Dai, Zhenhua Zhu, Tianyu Fu, Chiyue Wei, Bangyan Wang, **Xiangyu Li**, Yuan Xie, Huazhong Yang, and Yu Wang. 2022. DIMMining: Pruning-Efficient and Parallel Graph Mining on Near-Memory-Computing. In *Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA 2022).*
ISCA 2022 (CCF-A, THCPL-A). *Paper: ACM DL.*