

韦临烨

✉ lywei25@stu.pku.edu.cn | ☎ 13770579196 | 🌐 <https://linyewei526.github.io>

📍 北京市 | 🏠 谷歌学术

教育经历

北京大学 集成电路学院 直博

2025 年 9 月 – 至今

2025 级博士二班班长、研究生会就业实践工作组副组长

东南大学 吴健雄学院 (荣誉) / 未来技术学院 本科

2021 年 9 月 – 2025 年 6 月

未来技术学院一班团支部书记

研究兴趣

- **高效大模型推理**: 探索 LLM / dLLM / MoE 的推理加速、解码优化与稀疏执行
 - 推测解码加速: 研究 speculative decoding 在不同应用场景、硬件设备和解码范式下的优化, 提高草稿接收效率, 降低验证开销 [1][2][4]
 - Diffusion Large Language Models (dLLMs) 推理优化: 针对此类采用双向注意力的新模型和解码范式, 重构推理过程中迭代解掩码、cache 管理等关键执行步骤 [1][2]
 - 混合专家模型高效解码: 研究 MoE 模型推理过程中专家路由的动态变化特征, 以实现更稀疏的专家激活及其与新解码范式更高效的结合 [1]
- **算法-系统协同设计**: 探索模型计算、访存行为与系统执行之间的协同优化加速
 - AI 系统性能建模: 分析 AI 系统在硬件设备实际执行时的计算 / 访存行为, 结合 roofline model、运算强度讨论软硬件协同时的延迟瓶颈 [1][2]
 - 面向部署的系统优化: 考虑 AI 系统的执行 pipeline 向硬件设备的映射, 进行面向真实部署的硬件友好型优化, 包括均衡负载、压缩冗余运算、减少单元停滞等 [3]
 - 低开销免训练优化范式: 重点关注 plug-and-play 加速方案, 实现与现有推理框架和通用计算范式的可兼容性与快速迁移能力
- **多模态生成与高效计算**: 探索语音、视觉与 3D 场景中的高效生成与实时计算
 - 生成式语音识别: 针对基于 LLM 的自动语音识别 (ASR) 模型在转录精度和效率上的 trade-off, 研究解决 LLM 高解码延迟与 ASR 系统实时转录需求的矛盾, 实现端到端推理效率提升 [4]
 - 实时 3D 渲染优化: 分析 3D Gaussian Splatting 算法在三维渲染过程中的高斯基元冗余与硬件负载不均衡, 提供在算法优化和 CUDA 执行层面的解决方案 [3]
 - 边端图像生成: 面向 AI4S 背景下的人工数据集构造问题, 研究 diffusion 图像生成模型在边端系统中的可落地性 (DAC 2025 Contest)

论文工作

(* 表示贡献相同)

[1] **[ICML 2026, CCF-A]** TEAM: Temporal-Spatial Consistency Guided Expert Activation for MoE Diffusion Language Model Acceleration

LinYE Wei, Zixiang Luo, Pingzhi Tang, Meng Li

[2] **[DAC 2026, CCF-A]** Orchestrating Dual-Boundaries: An Arithmetic Intensity Inspired Acceleration Framework for Diffusion Language Models

LinYE Wei, Wenjue Chen, Pingzhi Tang, Xiaotian Guo, Le Ye, Runsheng Wang, Meng Li

[3] **[DAC 2025, CCF-A]** SpecASR: Accelerating LLM-based Automatic Speech Recognition via Speculative Decoding

LinYE Wei, Shuzhang Zhong, Songqiang Xu, Runsheng Wang, Ru Huang, Meng Li

[4] **[ICCAD 2025, CCF-B]** No Redundancy, No Stall: Lightweight Streaming 3D Gaussian Splatting for Real-time Rendering

LinYE Wei, Jiajun Tang, Fan Fei, Boxin Shi, Runsheng Wang, Meng Li

[5] **[WF-IoT 2025]** VR-YOLO: Enhancing PCB Defect Detection with Viewpoint Robustness Based on YOLO

Hengyi Zhu*, LinYE Wei*, He Li

发明专利

一种用于自动语音识别推测解码的草稿序列复用方法 (学生第一发明人)	2025
一种用于自动语音识别推测解码的两阶段稀疏树预测方法 (学生第一发明人)	2025

荣誉激励

北京大学博士研究生校长奖学金 (2/123)	2025
北京大学人工智能研究院博士研究生院长奖学金 (全院 20 人)	2025
东南大学未来技术“太湖”奖学金	2023
东南大学优秀共青团员	2023
东南大学“至善学子”奖学金	2022
东南大学校长奖学金 (6/242)	2022
东南大学三好学生	2022 / 2023 / 2024

竞赛获奖

DAC 2025 多模态生成系统设计竞赛 (CCF-A 会议主办) 全球第一名	2025
全国大学生电子设计竞赛 全国二等奖	2023
全国大学生电子设计竞赛 江苏赛区一等奖	2023
全国大学生数学竞赛 全国一等奖	2023
江苏省高等数学竞赛 二等奖	2022
全国大学生数学建模竞赛 江苏赛区一等奖	2022