雷浩东

个人主页：Homepage,GitHub,Google scholar
研究方向：AIGC、Speculative Sampling、AI Agent Memory
电话/微信：19521342247 | 邮箱：leihaodong@seu.edu.cn | 现居：江苏南京

教育背景

东南大学 **软件工程** **硕士** 2024.09-2027.06

- **PAttern Learning and Mining Lab 成员**
- **研究方向：**大模型投机采样加速、AI 智能体记忆

海南大学 **计算机科学与技术** **学士** 2020.09-2024.06

- **GPA：**3.88/4.0, **均分: 90.96**
- **综合排名：**3/123(2.44%)

论文(*指共同作者、 †指通讯作者)

- **H Lei**, D Wang*†, H Wang*†, et al.“MemCoT: Test-Time Scaling through Memory-Driven Chain-of-Thought”
  First Author, Submitted to ACMMM 26
- **H Lei**, H Wang†, et al. "Parallel-Path Relaxed Speculative Jacobi Decoding for Accelerating Auto-Regressive Text-to-Image Generation"
  First Author, Submitted to ECCV 26
- **H Lei**, H Wang†, et al. "Fast Inference of Visual Autoregressive Model with Adjacency-Adaptive Dynamical Draft Trees"
  First Author, Submitted to IJCV 26
- X Tan*, W Weng*, **H Lei**, et al. "EasyTune: Efficient Step-Aware Fine-Tuning for Diffusion-Based Motion Generation."
  Third Author, ICLR26

实习经历

1、科研任务部-前沿探索中心-算法实习生 上海人工智能实验室 2025.10–至今
上海

- **AI Agent 长文本对话记忆：**设计 MemCoT 测试时多尺度记忆系统。针对 LoCoMo 数据集 300 + 轮、9K+tokens 跨月超长对话场景，传统 RAG 语义稀释与上下文碎片化问题，提出 MemCoT 思考-记忆自进化框架。设计多尺度长时记忆感知模块，融合“细粒度检索+粗粒度观察”大幅提升全局结构完整性；引入语义与情景轨迹双短时记忆，支持智能体基于历史失败轨迹动态分解和修剪复杂查询。在 LoCoMo 数据集中取得 SOTA，GPT-4o-mini 下较纯 RAG 基线 F1 从 42.25 提升至 59.83 。以第一作者投稿至 ACMMM26 。

科研经历

1、图像自回归大模型推理加速 东南大学&新加坡管理大学 2025.07-2026.03

- **投机采样草稿树推广：**针对自回归文本生成图像模型推理速度慢的问题，引入多序列草稿树结构，将传统的 SJD 的链式依赖推广为层次化树状搜索，显著提升了 Draft Token 的接受率。
- **图像生成放宽加速：**提出跨序列松弛采样机制（Cross-Relaxed Strategy），利用不同序列间的语义相似性进一步优化放宽采样分布。在 Parti-Prompts、MSCOCO2017 等标准数据集上验证，该方法实现了约 4.14x 至 4.18x 的加速比，性能优于前沿方法。
- **成果：** Parallel-Path Relaxed Speculative Jacobi Decoding for Accelerating Auto-Regressive Text-to-Image Generation, 投稿至 ECCV26, 一作。

2、PALM 与 SMU 多模态大模型联合项目 东南大学&新加坡管理大学 2024.07-2025.07

- **图像生成区域相似性加速研究：**提出树状 Speculative sampling 方法下相邻草稿树的相似性问题. 根据不同像素块对应的草稿树深度与宽度不同，动态调整草稿树的形状。设计 Visual AR model 的 Speculative sampling 动态草稿树方法(ADT-Tree)， 使 Anole 模型在 MS-COCO 上的接受长度达到 3.4，与基线模型相比，推理加速了 2.21 倍。
- **成果：** Fast Inference of Visual Autoregressive Model with Adjacency-Adaptive Dynamical Draft Trees 投稿至 IJCV26, 一作

3、基于文本的动作生成 东南大学&新加坡管理大学 2024.12-2025.11

- **扩散模型后训练**: 提出一种用于扩散模型的微调框架 EasyTune，通过解耦递归依赖关系实现：（1）密集高效优化；（2）内存高效训练；（3）细粒度对齐。
- **成果：** EasyTune: Efficient Step-Aware Fine-Tuning for Diffusion-Based Motion Generation.(ICLR26, 三作)

实践技能

- 熟悉常见大模型轻量化算法，包括但不限于蒸馏、剪枝、量化、投机采样
- 熟悉常见 AI Agent Memory 构建机制与检索机制，包括但不限于 Rag、知识图谱、Multi-Agent
- 熟悉常见 AIGC 生成范式，如 DDPM、DDIM、CFG

竞赛经历

2023 海峡两岸暨港澳地区大学生计算机创新作品赛 全国一等奖	2023.06
第 25 届中国机器人及人工智能大赛人工智能创意赛 国家级三等奖	2023.06
2023 中国高校计算机大赛团体程序设计天梯赛“珠峰争鼎” 省级一等奖	2023.05
第 13 届全国大学生数学竞赛大赛 省级一等奖	2021.10