Haodong Lei

Website: Homepage, GitHub, Google scholar
Research Interests: AIGC、Speculative Sampling、AI Agent Memory
Phone/WeChat:19521342247 | Email: leihaodong@seu.edu.cn

Education

Southeast University | Software Engineering | Master | Sep 2024 - Jun 2027

- Member of Pattern Learning and Mining (PALM) Lab
- Research Focus: Large Language Model Speculative Sampling Acceleration, AI Agent Memory

Hainan University | Computer Science and Technology | Bachelor | Sep 2020 - Jun 2024

- **GPA:** 3.88/4.0, **Scores: 90.96**
- Overall Ranking: 3/123 (2.44%)

Publications (*denotes equal contribution, †denotes corresponding authors)

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

Internship Experience

Shanghai AI Lab | Shanghai AI Lab | Oct 2025 - Present | Shanghai

- **AI Agent Long-Context Dialogue Memory:** Designed MemCoT, a multi-scale memory system for test-time inference. Addressed challenges in LoCoMo dataset with 300+ turns, 9K+ tokens cross-month dialogues where traditional RAG suffers from semantic dilution and context fragmentation. Proposed Memory-Thought self-evolution framework with multi-scale long-term memory perception, integrating fine-grained retrieval and coarse-grained observation. Introduced dual short-term memory (semantic and situational trajectories) enabling agents to dynamically decompose and prune complex queries based on historical failure traces. Achieved SOTA on LoCoMo, improving F1 from 42.25 to 58.08 with GPT-4o-mini compared to pure RAG baseline. Submitted as first author to ACM MM 26.

Research Experience

1、Image Autoregressive LLM Inference Acceleration | Southeast University & Singapore Management University | Jul 2025 - Mar 2026

- **Extended speculative sampling draft tree:** Introduced multi-sequence draft tree structure for slow autoregressive text-to-image generation, expanding traditional SJD chain dependencies to hierarchical tree search, significantly improving Draft Token acceptance rate.
- **Relaxed sampling for image generation:** Proposed Cross-Relaxed Strategy leveraging semantic similarity between sequences to optimize relaxed sampling distribution. Achieved

approximately 4.14x to 4.18x speedup on Parti-Prompts and MSCOCO2017 datasets.

- **Publication:** Parallel-Path Relaxed Speculative Jacobi Decoding for Accelerating Auto-Regressive Text-to-Image Generation, submitted to ECCV 26, First Author.

2、PALM & SMU Multimodal LLM Joint Project — Southeast University & Singapore Management University — Apr 2025 - Jul 2025

- **Similarity-based acceleration for image generation:** Proposed tree-based Speculative Sampling addressing similarity issues in adjacent draft trees. Dynamically adjusted draft tree shape based on different pixel block depths and widths. Designed Adaptive Draft Tree (ADT-Tree) method for Visual AR model Speculative Sampling, enabling Anole model to achieve acceptance length of 3.4 on MS-COCO with 2.21x inference speedup compared to baseline.
- **Publication:** Fast Inference of Visual Autoregressive Model with Adjacency-Adaptive Dynamical Draft Trees, submitted to IJCV 26, First Author.

3、Text-Based Motion Generation — Southeast University & Singapore Management University — Dec 2024 - Nov 2025

- **Diffusion model post-training:** Proposed EasyTune, a fine-tuning framework for diffusion models achieving (1) dense efficient optimization, (2) memory-efficient training, and (3) fine-grained alignment by decoupling recursive dependencies.
- **Publication:** EasyTune: Efficient Step-Aware Fine-Tuning for Diffusion-Based Motion Generation, accepted at ICLR 26, Third Author.

Skills

- Large model lightweight algorithms: distillation, pruning, quantization, speculative sampling
- AI Agent Memory construction and retrieval mechanisms: RAG, knowledge graphs, Multi-Agent systems
- Common AIGC generation paradigms: DDPM, DDIM, CFG

Awards

First Prize, 2023 Cross-Strait Computer Innovation Competition, National Level — 2023.06

Third Prize, 25th China Robot and AI Competition - AI Creative Track, National Level — 2023.06