

LinYE Wei

✉ lywei25@stu.pku.edu.cn | ☎ +86 13770579196 | 🌐 <https://linyewei526.github.io>

📍 Beijing, China | 📄 Google Scholar

Education

Peking University School of Integrated Circuits Ph.D. Sep. 2025 – Present

Class Monitor of Ph.D. Class 2; Deputy Head of Career Practice Group, Graduate Student Union

Southeast University Chien-Shiung Wu College B.S. Sep. 2021 – Jun. 2025

Secretary of Youth League Branch, Class 1, Future Technology Institute

Research Interests

- **Efficient LLM Inference:** Exploring inference acceleration, decoding optimization, and sparse execution for LLMs, dLLMs, and MoE models
 - Speculative decoding acceleration: studying speculative decoding across application scenarios, hardware platforms, and decoding paradigms to improve draft acceptance and reduce verification overhead [1][2][4]
 - Diffusion Large Language Model (dLLM) inference: redesigning key execution steps such as iterative unmasking and cache management for bidirectional-attention dLLMs [1][2]
 - Efficient MoE decoding: studying dynamic expert routing in MoE inference to enable sparser expert activation and more efficient integration with new decoding paradigms [1]
- **Algorithm-System Co-design:** Exploring co-optimization across model computation, memory access, and system execution
 - AI system performance modeling: analyzing compute and memory behaviors of AI systems on real hardware, with roofline modeling and arithmetic intensity for latency bottleneck analysis [1][2]
 - Deployment-oriented system optimization: developing hardware-friendly optimizations for deployment pipelines, including load balancing, redundancy reduction, and stall mitigation [3]
 - Low-overhead training-free optimization: focusing on plug-and-play acceleration schemes with compatibility and fast migration across existing inference frameworks
- **Multimodal Generation and Efficient Computing:** Exploring efficient generation and real-time computation for speech, vision, and 3D workloads
 - Generative ASR: addressing the trade-off between transcription quality and efficiency in LLM-based automatic speech recognition (ASR), with a focus on end-to-end inference acceleration [4]
 - Real-time 3D rendering optimization: analyzing Gaussian primitive redundancy and hardware load imbalance in 3D Gaussian Splatting, with algorithmic and CUDA-level solutions [3]
 - Edge-side image generation: studying the deployability of image generation diffusion models on edge systems for synthetic dataset construction in AI4S settings (DAC 2025 Contest)

Publications

(* indicates equal contribution)

[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

Patents

A Draft Sequence Recycling Strategy for Speculative Decoding in ASR (First Student Inventor)	2025
A Two-pass Sparse-tree Prediction Method for Speculative Decoding in ASR (First Student Inventor)	2025

Honors and Awards

Presidential Scholarship for Ph.D. Students, Peking University (2/123)	2025
Dean's Scholarship for Ph.D. Students, Institute for AI, Peking University (20 recipients institute-wide)	2025
Future Technology TAIHU Scholarship, Southeast University	2023
Outstanding Youth League Member, Southeast University	2023
ZHISHAN Scholarship, Southeast University	2022
Presidential Scholarship, Southeast University (6/242)	2022
Merit Student, Southeast University	2022 / 2023 / 2024

Competition Awards

DAC 2025 System Design Contest (Hosted by a CCF-A Conference) 1st Place Worldwide	2025
National Undergraduate Electronic Design Competition National Second Prize	2023
National Undergraduate Electronic Design Competition Jiangsu Provincial First Prize	2023
National College Student Mathematics Competition National First Prize	2023
Jiangsu Province Advanced Mathematics Competition Second Prize	2022
National Undergraduate Mathematical Contest in Modeling Jiangsu Provincial First Prize	2022