

裴旻楠

(+86) 152-2930-7067 | peiminnan19@mails.ucas.ac.cn | Homepage



教育经历

2023 至今	中国科学院大学 · 自动化研究所（直博） · GPA: 3.89/4.0 · Top 10% 导师：程健、李钢；人工智能学院学生会主席（2024）	北京
2019–2023	中国科学院大学 · 电子信息工程（本科） · GPA: 3.93/4.0 · Top 5% 2022 春季公派访学：UC Berkeley · EECS · GPA: 4.0/4.0	北京 伯克利

研究方向

大模型推理与智能体系统：LLM/VLA 端侧异构部署、双系统并行加速与 Agent Router 调度优化
3DGS 高效渲染：专用加速数据流及体系结构、渲染数据流与移动端实时渲染优化

研究成果 (* Equal Contribution)

已发表及已接收

Minnan Pei, Gang Li, Junwen Si, Zeyu Zhu, Zitao Mo, Peisong Wang, Zhuoran Song, Xiaoyao Liang, Jian Cheng. *GCC: A 3DGS Inference Architecture with Gaussian-Wise and Cross-Stage Conditional Processing*. **MICRO 2025**

Minnan Pei, Gang Li, Zeyu Zhu, Siting Wang, Junwen Si, Yu Feng, Fangxin Liu, Zhuoran Song, Xiaoyao Liang, Jian Cheng. *DeGS: A Scalable 3DGS Architecture via Decoupled Workload Parsing and Reorganization*. **MICRO 2026**

Siting Wang, Minnan Pei, Luoyang Sun, Cheng Deng, Kun Shao, Zheng Tian, Haifeng Zhang, Jun Wang. *SpatialViz-Bench: Automatically Generated Spatial Visualization Reasoning Tasks for MLLMs*. **ICLR 2026**

Zhuoran Song, Haozhe Jiang, Chunyu Qi, Minnan Pei, Gang Li, Xiaoyao Liang, Haibing Guan. *VQVLA: Motion-Aware Vector Quantization with Computation Fusion for VLA Acceleration*. **MICRO 2026**

Zeyu Zhu, Gang Li, Minnan Pei, Zitao Mo, Peihuan Ni, Peisong Wang, Tielong Liu, Jian Cheng. *KL-MoE: A Hierarchical MoE Pruning Framework Exploiting KL Divergence*. **DAC 2026**

Yuanhui Wang, Kunlong Liu, Minnan Pei, Zhangming Li, Peisong Wang, Qinghao Hu. *MemeBQ: Memory Efficient Binary Quantization of LLMs*. **AAAI 2026**

投稿中

Minnan Pei, Qiwei Dong, Yihan Zhou, Yuchen Zhu, ..., Gang Li, Jian Cheng. *From Splats to Silicon: A Workload-Centric Survey of Hardware-Efficient 3D Gaussian Splatting*. **IEEE TPDS**

Minnan Pei, Gang Li, Junwen Si, Yudong Mu, Zhihua Fan, Zeyu Zhu, Zitao Mo, Peisong Wang, Zhuoran Song, Xiaoyao Liang, and Jian Chen. *GRADE: Gaussian-Wise Rendering with Adaptive Dependency-Ordered Execution for Static and Dynamic 3DGS*. **IEEE TCAD**

Yudong Mu, Minnan Pei*, Zhihua Fan, Siting Wang, Weifan Guan, Gang Li, Jian Cheng, Zhuoran Song, Xiaoyao Liang, Dongrui Fan, Xiaochun Ye, Wenming Li. *DANCE: Dual Budgeting with Action-Driven Near-Memory Compaction for Flow-Matching VLA*. **HPCA 2027**

Yudong Mu, Minnan Pei*, Zhihua Fan, Zirui Ma, Gang Li, Zhuoran Song, Yu Huang, Dongrui Fan, Xiaochun Ye, Wenming Li. *HARP: A Hierarchical Attention and Representation Sparsity Accelerator for Diffusion Transformers*. **HPCA 2027**

Runhao Liu, Minnan Pei*, Fei Ding, Guangzhen Yao, You Li, Peng Xiao, Gang Li, Peng Zhang. *LATTICE: Constraint-Directed Scheduling, Memory Planning, and Pipeline Refinement for NPUs*. **HPCA 2027**

Siting Wang, Minnan Pei, Xiaofeng Wang, Zheng Zhu, Xinyu Cui, Cheng Deng, Jian Zhao, Guan Huang, Haifeng Zhang, Jun Wang. *π -StepNFT: Wider Space Needs Finer Steps in Online RL for Flow-based VLAs*. **NeurIPS 2026**

实习经历

理想汽车 基座模型-VLA 模型架构组2026.05-2026.09

• 完成 VLA 模型端侧推理加速，探索优化双系统 VLA 并行推理执行效率。

研究经历

3DGS 高效渲染：研究 3DGS 数据流、专用体系结构与移动端实时渲染优化。

• 提出面向高斯粒度的专用数据流，在专用硬件上实现 3 倍以上的渲染性能与能效提升，并支持设计空间探索和动态场景适配。

• 重构渲染数据流并设计可扩展的专用推理架构，在更大规模的计算阵列和多种分辨率下保持 80% 以上的阵列利用率。

• 将推理数据流嵌入 Vulkan 渲染框架，在华为、三星等主流手机上实现 3DGS 实时渲染 1.5 倍加速，并显著降低功耗。

大模型推理加速：面向 LLM 与 VLA 的高效推理和部署，开展算法、系统及专用硬件协同优化研究。

• 面向资源受限场景，实现多类大语言模型在 FPGA 平台上的低比特稳定推理。

• 研究模型量化、MoE 剪枝与推测解码，优化大模型推理效率。

• 结合近存计算与协同调度，设计专用 VLA 加速架构，相较于 SOTA 边缘 GPU 实现 10 倍加速。

• 针对边缘 GPU 的 VLA 访存瓶颈，利用双系统异步执行实现推理解耦与端侧加速。

LLM Agent 系统与应用：开展自动化芯片设计、异构部署、多模型协作及评测研究。

• 构建面向异构 FPGA 的自动化部署框架，打通模型解析、量化、算子映射和硬件生成的端到端工具链。

• 构建面向大语言模型 Agent 的自动化芯片设计与评测基准。

• 探索低成本、高质量的 token 生成方法，并评估多模型 Agent Router 的路由能力。

荣誉奖项

中国科学院大学三好学生	2020-2026
中国科学院大学优秀共青团员	2020-2023
中国科学院大学优秀学生干部	2020-2026
中国科学院大学优秀共青团干部	2020-2023
中国科学院大学一等学业奖学金	2020、2021、2022
北京市三好学生	2021
国家奖学金	2022
中国科学院大学大学生奖学金	2022
Robomaster 机甲大师赛二等奖	2022
中国科学院大学三好学生标兵	2022、2026
北京市优秀毕业生	2023

学术经历

论文审稿：ISCA · HPCA · CoRL · ACM MM · TCAD · TCAS-I · TVLSI

学术汇报：MICRO 2025 Oral · MICRO 2026 Oral · CCF Sys 2025 大会讲者

学生助教：电路原理（中国科学院大学 2022 年秋季学期） · 国科大 2022 级人工智能本科生朋辈导师

专业技能

编程语言	Python, C/C++, CUDA, Verilog, MATLAB, Chisel, HTML
框架工具	PyTorch, Git, Linux, Docker, CACTI, Ramulator, Vulkan, Codex, Claude Code
技能特长	CET-4, CET-6, TOEFL, 篮球, 声乐, 二胡