倪培洹

北京市海淀区中关村东路 95 号，中国科学院自动化研究所

+86 158 0375 7631 ◇ nipeihuan24@mails.ucas.ac.cn ◇ peihuanni.github.io

研究方向

面向大模型训练与推理的人工智能基础设施和分布式系统；大语言模型与视觉–语言–动作模型的高效推理架构；硬件加速器与具身人工智能。

教育经历

吉林大学，学士 2020 年 9 月 – 2024 年 6 月

电子科学与工程学院，微电子科学与工程 导师：阎凯歌

CET-4：603，CET-6：553

中国科学院大学，硕士研究生 2024 年 9 月 – 2027 年 6 月

人工智能学院，人工智能 导师：程健、李钢

中国科学院自动化研究所，智能计算与学习实验室（CLab）

GPA：3.81

科研经历

第一作者论文

Peihuan Ni, Zitao Mo, Tielong Liu, Hongli Wen, Zeyu Zhu, Minnan Pei, Junwen Si, Weifan Guan, Peisong Wang, Qinghao Hu, Gang Li and Jian Cheng. **APEX: Integer-only Non-linear Function Approximation for Efficient Cross-Modal Inference.** In **DATE 2026**. **(CCF-B，计算机体系结构顶级会议)**

- 发现不同模态模型在量化后的数据分布存在显著差异，使其对非线性函数近似误差表现出不同敏感性，而统一保留高精度会导致计算位宽和硬件开销过高。
- 提出统一的整数非线性计算架构，融合多种常用非线性算子的计算图，并通过比特级剪枝减少计算位宽，同时保持模型精度。
- 在语言模型和视觉模型上分别提升最高约 0.7% 和 1.3% 的任务精度，并在多模态模型上实现近乎无损推理。
- 相较已有方法，面积降低 1.73–8.71×，功耗降低 1.21–10.83×。

Peihuan Ni, Zeyu Zhu, Weifan Guan, Gang Li, Zitao Mo, Tielong Liu, Junwen Si, Qinghao Hu, Peisong Wang and Jian Cheng. **ACE-VLA: Action-Centric Evaluation for Token Pruning in Vision-Language-Action Models.** In **CoRL 2026**. **(智源 A 类会议，具身智能顶级会议)**

- 指出现有 VLA 剪枝方法主要沿用 VLM 中的文本–视觉注意力评价方式，未充分利用动作输出中包含的任务阶段和运动语义信息。
- 提出以动作输出为中心的梯度评价指标，并使用基于运动轨迹的动态 Token 池维护策略降低重要性评估开销。
- 从运动学角度建模机器人的运动趋势，而非仅依据瞬时运动状态，实现动态剪枝比例调整。
- 在 $\pi_{0.5}$ 和 π_0 模型上分别提升约 1.4% 和 2.2% 的任务成功率，推理速度提高 1.33–1.57×，并在 OpenVLA-OFT 上取得优于现有基线的综合性能。
- 通过真机实验验证方法在真实场景中的可迁移性，并提升 VLA 剪枝过程的可解释性。

共同作者论文

Zeyu Zhu, Gang Li, Minnan Pei, Zitao Mo, **Peihuan Ni**, Peisong Wang, Tielong Liu and Jian Cheng. **KL-MoE: A Hierarchical MoE Pruning Framework Exploiting KL Divergence.** In **DAC 2026**. **(CCF-A，并行与分布计算顶级会议)**

Tielong Liu, Zeyu Zhu, Gang Li, Zitao Mo, Chen BAI, Junwen Si, **Peihuan Ni**, Minnan Pei, Peisong Wang, Zhuoran Song, Xiaoyao Liang and Jian Cheng. **FLUTE: A LUT-Based Floating-Point Engine for Efficient LLM Acceleration Exploiting Core-Tail Hierarchical Alignment.** In **MICRO 2026**. **(CCF-A，计算机体系结构顶级会议)**

Junwen Si, Gang Li, Tielong Liu, **Peihuan Ni**, Minnan Pei, Zeyu Zhu, Zitao Mo, Peisong Wang, Zhongfeng Wang and Jian Cheng. **DOMI: A Dataflow Optimization Framework for Efficient Mamba Inference on FPGAs.** Submitted to **TCAD**. **(CCF-A，计算机体系结构顶级期刊)**

Zeyu Zhu, Gang Li, Peisong Wang, **Peihuan Ni**, Tielong Liu, Minnan Pei, Junwen Si, Zhuoran Song, Xiaoyao Liang and Jian Cheng. **DALI: A Workload-Aware Offloading Framework for Efficient MoE Inference on Local PCs.** Submitted to **ATC 2026**. **(CCF-A，并行与分布计算顶级会议)**

实习经历

华为，ICT BG **Ascend NPU 低比特 FlashAttention 软硬协同推理优化** **2026 年 7 月 – 2027 年 7 月**

- 面向矩阵与非矩阵算力不对称扩张，针对低比特 FlashAttention 开展 Ascend NPU 上的软硬协同推理优化。

项目经历

VLA 模型量化压缩与 FPGA 端侧部署 项目负责人

- 设计 W4A8 量化方案，其中权重采用 4-bit 量化，激活值和 KV Cache 采用 8-bit 量化，并结合纯整数非线性近似实现近乎无损的端侧推理。
- 实现基于 FP16–INT8 MAC 的量化与反量化融合算子，减少中间数据回写和片外访存，提高激活值的片上驻留比例。
- 复用 FP16–INT8 MAC 完成 RoPE 等算子的计算，并调整部分计算顺序，提高计算单元和片上数据的复用效率。
- 面向 W4A8 混合精度矩阵乘法设计比特串行计算单元，以及支持不同注意力头维度的 KV 重排序模块和 BRAM 到计算阵列的 Crossbar 数据映射。
- 设计统一的纯整数非线性计算单元，在降低硬件面积和功耗的同时支持多类非线性算子，相关成果发表于 DATE 2026。
- 参与 VLIW 编译器设计，负责片上数据流、指令处理器和算子控制模块；通过地址偏移和多种偏移模式，将单次模型推理所需指令数降低 32× 以上。
- 在 GPU 上实现 Bit-Accurate 仿真器，并完成与 Vivado 模块级仿真及 FPGA 整网运行结果的逐级对齐。
- 完成 DMA 访存通路调试、HBM 带宽优化、VHK158 开发板数据通路调试、时序收敛及 PetaLinux 系统制作，实现模型在 FPGA 上的完整部署。

Ascend A5 NPU 低比特 FlashAttention 推理优化（华为实习项目） 项目负责人

- 面向矩阵与非矩阵算力不对称扩张，开展 Ascend A5 NPU 上的低比特 FlashAttention 软硬协同推理优化。
- 通过 Direct-P 使用低成本线性运算替代 exp 计算与量化，大幅减少 Vector 侧计算。
- 重构 O/l 更新算法，利用单条 MMAD 指令完成 O/l 更新，减少 Vector 计算与中间数据搬运。
- 采用转置累加降低 Nz–ND 数据排布转换开销，缩短关键路径。

奖项荣誉

综合荣誉

中国科学院大学三好学生（2025）

竞赛获奖

英特尔杯全国大学生电子设计竞赛邀请赛三等奖（2022）

全国大学生数学建模竞赛二等奖（2022）

中国大学生数学竞赛三等奖（2023）

社会服务

WACV、EMNLP、CoRL 审稿人