PEIHUAN NI

+86 158 0375 7631 ◇ nipeihuan24@mails.ucas.ac.cn ◇ peihuanni.github.io
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Haidian District, Beijing, China

RESEARCH INTERESTS

AI infrastructure and distributed systems for large-model training and inference; efficient inference architectures for large language models and vision-language-action models; hardware accelerators and embodied artificial intelligence.

EDUCATION

University of Chinese Academy of Sciences Sep. 2024 – Jun. 2027 (expected)
Master's student in Artificial Intelligence, School of Artificial Intelligence
Institute of Automation, Chinese Academy of Sciences; Efficient Intelligent Computing and Learning Lab (CLab)
Advisors: Jian Cheng and Gang Li GPA: 3.81

Jilin University Sep. 2020 – Jun. 2024
Bachelor's degree in Microelectronics Science and Engineering
College of Electronic Science and Engineering Advisor: Kaige Yan
College English Test: CET-4: 603; CET-6: 553

RESEARCH EXPERIENCE

First-Author Publications

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)

- Identified substantial differences in quantized data distributions across model modalities, leading to varying sensitivity to non-linear approximation errors; retaining high precision uniformly incurs excessive computational bit widths and hardware costs.
- Proposed a unified integer-only architecture that merges the computation graphs of common nonlinear operators and uses bit-level pruning to reduce bit widths while preserving model accuracy.
- Improved task accuracy by up to approximately 0.7% for language models and 1.3% for vision models, while enabling near-lossless inference for multimodal models.
- Achieved area reduction factors of 1.73–8.71× and power reduction factors of 1.21–10.83× relative to existing methods.

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**. (Top Conference on Robot Learning)

- Identified that existing VLA pruning methods largely adopt text–vision attention metrics from VLMs, underusing task-phase and motion semantics encoded in action outputs.
- Proposed an action-centric gradient-based importance metric and a motion-trajectory-based dynamic token pool to reduce importance evaluation overhead.
- Modeled robot motion trends from a kinematic perspective to dynamically adjust pruning ratios beyond instantaneous motion states.
- Improved task success rates by approximately 1.4% on $\pi_{0.5}$ and 2.2% on π_0, with 1.33–1.57× inference speedups; outperformed existing baselines in overall performance on OpenVLA-OFT.
- Validated transferability through real-robot experiments and improved the interpretability of VLA token pruning.

Co-Authored Publications

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)

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)

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)

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)

INTERNSHIP EXPERIENCE

Huawei, ICT BG Jul. 2026 – Jul. 2027

Hardware–Software Co-Design for Low-Bit FlashAttention Inference on Ascend NPUs

- Optimize low-bit FlashAttention inference on Ascend NPUs through hardware–software co-design under asymmetric scaling of matrix and non-matrix compute throughput.

PROJECTS

VLA Model Quantization and FPGA Edge Deployment **Project Lead**

- Designed a W4A8 quantization scheme with 4-bit weights and 8-bit activations and KV cache, combined with integer-only nonlinear approximations for near-lossless edge inference.
- Implemented fused quantization and dequantization operators using FP16–INT8 multiply-accumulate (MAC) units, reducing intermediate writebacks and off-chip memory accesses while keeping more activations on chip.
- Reused FP16–INT8 MAC units for operators such as RoPE and reordered selected computations to improve compute-unit and on-chip data reuse.
- Designed bit-serial compute units for W4A8 mixed-precision matrix multiplication, a KV reordering module supporting different attention head dimensions, and crossbar-based data mapping from BRAM to the compute array.
- Designed a unified integer-only nonlinear unit supporting multiple operators with reduced hardware area and power consumption; results published at DATE 2026.
- Contributed to VLIW compiler design, with responsibility for on-chip dataflow, the instruction processor, and operator control modules; reduced the instruction count per model inference by over 32$\times$ using address offsets and multiple offset modes.
- Built a GPU-based bit-accurate simulator and verified consistency with Vivado module-level simulations and full-model execution on FPGA.
- Completed end-to-end FPGA deployment, including DMA memory-access debugging, HBM bandwidth optimization, VHK158 board data-path debugging, timing closure, and PetaLinux system builds.

Low-Bit FlashAttention Inference Optimization on the Ascend A5 NPU (Huawei Internship) **Project Lead**

- Optimized low-bit FlashAttention through hardware–software co-design under asymmetric scaling of matrix and non-matrix compute throughput.
- Applied Direct-P to replace exp evaluation and quantization with low-cost linear operations, substantially reducing Vector-side computation.
- Reworked the O/l update algorithm to complete the update with a single MMAD instruction, reducing Vector computation and data movement.
- Used transposed accumulation to reduce Nz–ND layout-conversion overhead and shorten the critical path.

HONORS AND AWARDS

Outstanding Student, University of Chinese Academy of Sciences (2025)
Third Prize, Intel Cup National Undergraduate Electronic Design Contest – Invitational (2022)
Second Prize, China Undergraduate Mathematical Contest in Modeling (2022)
Third Prize, Chinese Mathematics Competitions for College Students (2023)

ACADEMIC SERVICE

Reviewer for WACV, EMNLP, CoRL.