Weikang (Zachary) Meng

School of Computer Science and Technology | Harbin Institute of Technology, Shenzhen

E-mail | Homepage | Google Scholar | WeChat: MWK_0411 | Phone: +86 18804636326

EDUCATION

Harbin Institute of Technology, Shenzhen **Sep. 2023 – Present (Expected Dec. 2027)**

Ph.D. in Computer Science and Technology, School of Computer Science and Technology

- Advisor: Prof. Zheng Zhang
- Research Interests: Linear Attention, Efficient Foundation Models

Harbin Institute of Technology **Sep. 2019 – Jul. 2023**

B.S. in Information and Computational Science, School of Mathematics

- GPA: 85.98/100, Recommended for Direct Ph.D. Admission

SELECTED PUBLICATIONS & RESEARCH EXPERIENCE

PolaFormer++: Polarity-aware Channel-Spiky Linear Attention for Vision Transformers (TPAMI 2026, CCF-A TOP, IF: 20.8) [Link]

Authors: ***Weikang Meng***, Liangyu Huo, Yadan Luo, Yaowei Wang, Yingjian Li, Zheng Zhang, Heng Tao Shen

- Extended PolaFormer with channel-wise spikiness modeling to improve attention sharpness and feature discrimination, which demonstrated strong transferability across multiple vision tasks.

Norm Direction: Restoring the Missing Query Norm in Vision Linear Attention (ICML 2026, CCF-A TOP) [Link]

Authors: ***Weikang Meng***, Yadan Luo, Liangyu Huo, Yaowei Wang, Xin Li, Zheng Zhang

- Revealed the missing-query-norm in linear attention and derived Positive Sequence Entropy theory.
- Proposed norm-aware kernel functions and geometric similarity functions, forming a novel linear attention.

PolaFormer: Polarity-aware Linear Attention for Vision Transformers (ICLR 2025, CCF-A TOP, 100+ citations) [Link]

Authors: ***Weikang Meng***, Yadan Luo, Xin Li, Dongmei Jiang, Zheng Zhang

- Proposed polarity decomposition to explicitly model positive and negative Query-Key interactions.
- Established theoretical conditions for sharper attention distributions, achieving SOTA performance.

STILL: Selecting Tokens for Intra-Layer Hybrid Attention to Linearize LLMs (NeurIPS Under Review) [Link]

Authors: ***Weikang Meng***, Liangyu Huo, Yadan Luo, Jiawen Guan, Jingyi Zhang, Yingjian Li, Zheng Zhang

- Proposed intra-layer hybrid attention for linearizing LLM, with softmax attention and linear attention.
- Developed Self-Saliency Score for Token selection and Norm-Preserved Feature Map for linear branches, which improving long-context efficiency and restoring model capability.

Ascend-Aware Efficient Transformer Architecture Optimization (Huawei Joint Research Project, 2023.12–2024.12, RMB 1.2M in funding)

- Designed an Ascend-friendly linear-complexity attention to replace softmax attention, cutting per-step training time by 20% and memory by 15% on BERT/GPT2/LLaMA-scale models using mindspeed-llm.
- Honored with the HUAWEI Spark Award.

PROFESSIONAL SERVICE

Ascend-TLA-DeltaNet: The First Efficient Linear Attention Operator Library for Ascend NPU [Link]

- Built the first efficient linear attention operator library for Ascend NPUs; merged into the main repository of Huawei CANN (cann-recipes-train).
- Recognized as a *CANN Contributor* and *CANN Core Developer*.

Reviewer/Program Committee: NeurIPS, ICML, ICLR, ECCV, CVPR, AAAI, TIP, TCSVT, PR.

HONORS & AWARDS

- HUAWEI Spark Award: Ascend Hardware-Aware Transformer Operator Optimization.
- Honorable Mention, Mathematical Contest in Modeling (MCM/ICM) (2022)
- Provincial Second Prize, Chinese Mathematics Competition for College Students (2020)
- Provincial First Prize, National High School Mathematics Competition (2018)