WENXUAN ZENG

zwx.andy@stu.pku.edu.cn | 15388161786 | wenxuanzeng.netlify.app

Beijing | Blog: xuanland.blog.csdn.net | Google Scholar

Education

Peking University (PKU) September 2023 – Present
M.S., Microelectronics & Institute for Artificial Intelligence
Deputy Minister of Academic Science and Technology Innovation Department

University of Electronic Science and Technology of China (UESTC) September 2019 – June 2023
B.S., Software Engineering (GPA: 3.9 / 4.0)
Chairman of School Technology Studio and Leader of Machine Learning Group

Research Interests

The listed references mean the relevant works I have participated in.

- **LLM Optimization Algorithm**: Investigate sparse attention and efficient chain-of-thought (CoT) reasoning.
  - Sparse attention and KV cache compression for long-context LLM inference: investigate efficient KV cache optimizations, including re-computation, low-rank decomposition, block attention, static and dynamic token pruning [4][2][1].
  - Efficient CoT reasoning: investigate the current efficient CoT compression algorithms, including latent-space CoT, prompting-guided compression, and training-internalized CoT compression.
- **Efficient Model Compression/Acceleration**: Explore how to employ and optimize efficient compression algorithms for different models, including CNNs, ViTs, and LLMs.
  - Architecture design and neural architecture search: design and search efficient attention mechanisms [11], and automatically prune redundant non-linear layers such as ReLU and GeLU [10][3][11].
  - Knowledge distillation: employ logists-based distillation and feature-based distillation for tiny model performance enhancement [10][8][11][6].
  - Quantization: compress the activations and weights to low bit widths for efficient inference and explore the layer-wise precision allocation [8][6].
  - Efficient privacy-preserving deep learning: Investigate the deployment of private inference for CNNs, ViTs, and LLMs, and co-design both efficient AI algorithms (mentioned above) and inference protocols, enhancing the system efficiency and performance while preserving inference privacy for both user's data and proprietary model parameters.
- **Multimodal LLM (MLLM)**: Recent research focuses on how to give MLLM human thinking patterns, including multimodal chain-of-thought (MCoT) reasoning, which focuses on slow-thinking reasoning in MLLM.

Selected Publications

** indicates equal contribution, † indicates corresponding author.*

[1] UniCAIM: A Unified CAM/CIM Architecture with Static-Dynamic KV Cache Pruning for Efficient Long-Context LLM Inference
Weikai Xu*, **Wenxuan Zeng***, Qianqian Huang, Meng Li†, Ru Huang†
DAC 2025

[2] MPCache: MPC-Friendly KV Cache Eviction for Efficient Private Large Language Model Inference
Wenxuan Zeng, Ye Dong, Jinjin Zhou, Junming Ma, Jin Tan, Runsheng Wang, Meng Li†
Preprint 2025, arxiv.org/abs/2501.06807

[3] EQO: Exploring Ultra-Efficient Private Inference with Winograd-Based Protocol and Quantization Co-Optimization
Wenxuan Zeng, Tianshi Xu, Cheng Hong, Meng Li†, Runsheng Wang
Preprint 2024, https://arxiv.org/abs/2404.09404

[4] CoPriv: Network/Protocol Co-Optimization for Communication-Efficient Private Inference
Wenxuan Zeng, Meng Li†, Haichuan Yang, Wen-jie Lu, Runsheng Wang, Ru Huang
NeurIPS 2023

[5] MPCViT: Searching for Accurate and Efficient MPC-friendly Vision Transformer with Heterogeneous Attention
Wenxuan Zeng, Meng Li†, Wenjie Xiong, Tong Tong, Wen-jie Lu, Jin Tan, Runsheng Wang, Ru Huang
ICCV 2023

Experience

Institute for Artificial Intelligence, Peking University June 2022 – Present

- Role: Student researcher
- Advisor: Prof. Meng Li, Prof. Runsheng Wang
- Topics: Efficient AI algorithms and efficient private inference (mainly focus on protocol-algorithm co-optimization)
- Projects: Efficient attention architecture design [11], ReLU and GeLU pruning [10][11][3], efficient convolution optimization [8][10], KV cache compression for LLM inference [4][2][1], low-precision quantization [6][8], protocol-level optimizations for private inference [3][5][10][8][6]
- Industry collaboration: Record and launch the course about "Privacy-Preserving Machine Learning" with Ant Group

Knowledge Works Research Laboratory, Fudan University April 2022 – July 2022

- Role: Research intern
- Advisor: Prof. Yanghua Xiao, Dr. Jiangjie Chen
- Topics: Factuality and faithfulness in natural language generation, explainable AI
- Projects: Factual error correction [12]

Sichuan Key Laboratory of Network and Data Security, UESTC June 2021 – June 2022

- Role: Research intern
- Advisor: Prof. Fan Zhou, Dr. Zhiyuan Wang
- Topics: Graph neural networks, street-level IP geolocation, multivariate time series modeling
- Projects: Street-level IP geolocation with graph attention network [13]

Communication and Information Security Laboratory, Peking University August 2021 - May 2022

- Role: Research intern
- Advisor: Prof. Yuesheng Zhu, Dr. Wenyuan Yang
- Topics: Fine-grained image recognition and iris recognition

Honors and Achievements

Third Prize Scholarship at Peking University	2024
Merit Student Scholarship at Peking University	2024
Outstanding Graduate Student and Honors Research Certificate at UESTC	2023
Outstanding Student Scholarship at UESTC	2020 - 2022
"Innovation and Entrepreneurship Training Plan" at UESTC	2020 - 2022
Special Prize of "Tencent" Scholarship (3 Places at UESTC in Total)	2022
First Prize of "Shi Qiang" Scholarship (5 Places at UESTC in Total)	2021

Competition Awards

Badminton Competition of Institute for Artificial Intelligence at Peking University 2024
The 2nd Prize

Annual Symposium and Tech Day of Institute for Artificial Intelligence at Peking University 2023
Most Popular Award

China Collegiate Computing Contest 2023
National Special Prize (the 1st Place)

"Pan-Pearl River Delta" Collegiate Computer Work Competition 2022
National Third Prize

China Collegiate Computing Contest 2021
National First Prize (the 1st Place)

"Zhong Gong Cup" Sichuan Collegiate Computer Work Competition 2021
Provincial Special Prize

Publications

** indicates equal contribution, † indicates corresponding author.*

The publication list is available through Google Scholar (130+ citations): please click here.

[1] H^2EAL: Hybrid-Bonding Architecture with Hybrid Sparse Attention for Efficient Long-Context LLM Inference
Zizhuo Fu, Xiaotian Guo, **Wenxuan Zeng**, Shuzhang Zhong, Yadong Zhang, Peiyu Chen, Runsheng Wang, Le Ye, Meng Li†
Under Review

[2] UniCAIM: A Unified CAM/CIM Architecture with Static-Dynamic KV Cache Pruning for Efficient Long-Context LLM Inference
Weikai Xu*, **Wenxuan Zeng***, Qianqian Huang, Meng Li†, Ru Huang
DAC 2025

[3] OptiPrime: Efficient Private Inference at ImageNet Scale
Jiangrui Yu, Ye Yu, Si Chen, **Wenxuan Zeng**, Junfeng Fan, Runsheng Wang, Ru Huang, Meng Li†
Work in Progress

[4] MPCache: MPC-Friendly KV Cache Eviction for Efficient Private Large Language Model Inference
Wenxuan Zeng, Ye Dong, Jinjin Zhou, Junming Ma, Jin Tan, Runsheng Wang, Meng Li†
Preprint 2025, arxiv.org/abs/2501.06807

[5] FlexHE: A Flexible Kernel Generation Framework for Homomorphic Encryption-Based Private Inference
Jiangrui Yu, **Wenxuan Zeng**, Tianshi Xu, Renze Chen, Yun (Eric) Liang, Runsheng Wang, Ru Huang, Meng Li†
ICCAD 2024

[6] PrivQuant: Communication-Efficient Private Inference with Quantized Network/Protocol Co-Optimization
Tianshi Xu, Shuzhang Zhong, **Wenxuan Zeng**, Meng Li†, Runsheng Wang, Ru Huang
ICCAD 2024

[7] BAT: Behavior-Aware Human-Like Trajectory Prediction for Autonomous Driving
Haicheng Liao, Zhenning Li, Huanming Shen, **Wenxuan Zeng**, Dongping Liao, Guofa Li, Shengbo Eben Li, Chengzhong Xu†
AAAI 2024

[8] EQO: Exploring Ultra-Efficient Private Inference with Winograd-Based Protocol and Quantization Co-Optimization
Wenxuan Zeng, Tianshi Xu, Cheng Hong, Meng Li†, Runsheng Wang
Preprint 2024, https://arxiv.org/abs/2404.09404

[9] Kuaiji: the First Chinese Accounting Large Language Model
Jiayuan Luo, Songhua Yang, Xiaoling Qiu, Panyu Chen, Yufei Nai, **Wenxuan Zeng**, Wentao Zhang†, Xinke Jiang
Preprint 2024, https://arxiv.org/abs/2402.13866

[10] CoPriv: Network/Protocol Co-Optimization for Communication-Efficient Private Inference
Wenxuan Zeng, Meng Li†, Haichuan Yang, Wen-jie Lu, Runsheng Wang, Ru Huang
NeurIPS 2023

[11] MPCViT: Searching for Accurate and Efficient MPC-friendly Vision Transformer with Heterogeneous Attention
Wenxuan Zeng, Meng Li†, Wenjie Xiong, Tong Tong, Wen-jie Lu, Jin Tan, Runsheng Wang, Ru Huang
ICCV 2023

[12] Factual Error Correction via Iterative Constrained Editing
Jiangjie Chen*, Rui Xu*, **Wenxuan Zeng**, Changzhi Sun†, Lei Li, Yanghua Xiao†
AAAI 2023

[13] Connecting the Hosts: Street-Level IP Geolocation with Graph Neural Networks
Zhiyuan Wang*, Fan Zhou*†, **Wenxuan Zeng***, Goce Trajcevski, Chunjing Xiao, Yong Wang, Kai Chen
KDD 2022

Academic Service

- Serving as a reviewer of NeurIPS 2025