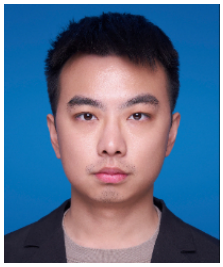
Zhaohui Xu

PhD Student in Low-Precision Inference and Efficient AI Systems
Shanghai, China | +86 17671341889 | xuzhh12025@shanghaitech.edu.cn
github.com/Zhaohui-Xu | ORCID 0009-0003-8334-6903



Education

ShanghaiTech University 2025—2028
PhD in Computer Science and Technology | Joint training with the Qizhi Institute | Supervisors: Prof. Yajun Ha (ShanghaiTech University); Prof. Li Jiang (Shanghai Jiao Tong University)

ShanghaiTech University 2022—2025
Master in Computer Science and Technology | Joint training with the Institute of Computing Technology, Chinese Academy of Sciences | Supervisors: Prof. Yajun Ha (ShanghaiTech University); Researcher Ying Wang (Institute of Computing Technology, Chinese Academy of Sciences)

China Three Gorges University 2018—2022
Bachelor in Computer Science and Technology

Research Focus & Technical Depth

I work on low-bit quantization, model compression, and efficient deployment for AI inference systems, with an emphasis on kernel-runtime co-optimization and heterogeneous software-hardware co-design.

Research areas: Low-bit quantization and model compression | Efficient inference and deployment | Kernel-runtime and heterogeneous-system co-optimization
Programming & systems: Python, C/C++, CUDA, Verilog; **low-precision inference:** INT8/W4A4 quantization, dynamic and mixed precision, pruning-aware optimization; **frameworks:** PyTorch custom operators, llama.cpp, vLLM, TensorRT; **kernels & profiling:** CUDA, Ascend C, GPU memory management, Nsight Systems, NVTX/XPU profiling.

Industry Experience

Baidu Kunlunxin | Deep Learning Framework Development Engineer Intern 2024-10 - 2025-01
Training Products Team
- **Technical focus:** PyTorch compatibility engineering, custom-operator integration, and performance-tooling extension for domestic XPU accelerators.
- Integrated and tested high-performance custom operators for PyTorch/XPU, adapted the NVTX backend for XPU, and implemented XSight System profiling visualization.

Huawei Technologies | AI Engineer Intern 2024-06 - 2024-09
2012 Laboratories — Ascend Solution Development Department
- **Technical focus:** FlashAttention operator tuning, workload adaptation, and fused implementation for Ascend NPUs.
- Tuned and adapted Ascend C FlashAttention and PromptFlashAttention fused operators, and evaluated accuracy and performance for speculative-inference workloads.

Selected Publications

1. COMET: Towards Practical W4A4KV4 LLMs Serving ASPLOS 2025
CCF A | Lian Liu, Long Cheng, Haimeng Ren, **Zhaohui Xu**, Yudong Pan, Mengdi Wang, Xiaowei Li, Yinhe Han, Ying Wang
- **Research role:** Research and implementation of the FMPQ fine-grained mixed-precision quantization method.
- Analyzed outlier distributions in LLM activations and developed FMPQ, a fine-grained mixed-precision quantization method that enables 4-bit activations and KV cache with modest accuracy loss.

2. Make LLM Inference Affordable to Everyone: Augmenting GPU Memory with NDP-DIMM HPCA 2025
CCF A | Lian Liu, Shixin Zhao, Bing Li, Haimeng Ren, **Zhaohui Xu**, Mengdi Wang, Xiaowei Li, Yinhe Han, Ying Wang
- **Research role:** Efficient hot/cold-neuron prediction using MLP pruning.
- Developed a heterogeneous inference approach that maps hot neurons to the GPU and offloads cold neurons to NDP-DIMMs, balancing compute efficiency with memory capacity.

3. Drift: Leveraging Distribution-based Dynamic Precision Quantization for Efficient Deep Neural Network Acceleration DAC 2024
CCF A | Lian Liu, **Zhaohui Xu**, Yintao He, Ying Wang, Huawei Li, Xiaowei Li, Yinhe Han
- **Research role:** Hardware-friendly dynamic mixed-precision quantization and the design and implementation of the dynamic quantization accelerator.
- Developed a model-agnostic, low-overhead dynamic mixed-precision quantization method that executes about 80% of INT8 data in INT4 with about 1% accuracy loss.

Awards & Recognition

- **2026** CANN Ecosystem Contributor Certificate — Core Developer
- **2025** Open Hardware 2025 Winner
- **2025** CCF Sys2025 Graph Computing System Design Competition Grand Prize (team member)