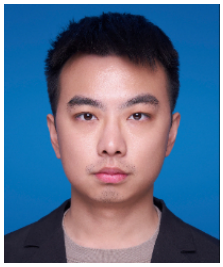

徐兆辉

低精度推理与高效 AI 系统方向博士生

上海，中国 · +86 17671341889 · xuzhh12025@shanghaitech.edu.cn
github.com/Zhaohui-Xu · ORCID 0009-0003-8334-6903

教育经历

上海科技大学 2025—2028 | 上海
计算机科学与技术博士 | 期智研究院联合培养 | 联合指导：哈亚军教授、蒋力教授

上海科技大学 2022—2025 | 上海
计算机科学与技术硕士 | 中国科学院计算技术研究所联合培养 | 联合指导：哈亚军教授、王颖研究员

三峡大学 2018—2022 | 宜昌
计算机科学与技术学士

研究方向与技术能力

面向 AI 推理系统开展低比特量化、模型压缩与高效部署研究，重点关注 Kernel/Runtime 协同优化及异构软硬件协同设计。

- **编程与系统：** Python、C/C++、CUDA、Verilog
- **低精度推理与压缩：** INT8/W4A4 量化、动态与混合精度推理、剪枝驱动的推理优化、内存优化
- **推理系统与框架：** PyTorch 自定义算子接入、llama.cpp、vLLM、TensorRT
- **算子与性能工程：** CUDA 与 Ascend C 算子开发、GPU 内存管理、Nsight Systems、NVTX/XPU Profiling

实习经历

百度昆仑芯 · 深度学习框架开发工程师（实习） 2024.10—2025.01
训练产品组

- **技术方向：** 面向国产 XPU 的 PyTorch 适配性开发、自定义算子接入与性能工具链扩展。
- 完成 moe_fc_fusion 等高性能算子的 PyTorch 自定义算子注册接入、算子测试与用户文档撰写。
- 围绕系统级性能优化场景调研 Nsight Systems，并形成分析材料。
- 完成 NVTX 的 XPU 后端适配，结合 XPU-Runtime 信息收集工具实现 XSight System 的 Profiling 可视化。

华为技术有限公司 · AI 工程师（实习） 2024.06—2024.09
2012实验室—昇腾解决方案开发部

- **技术方向：** 面向 Ascend NPU 的 FlashAttention 算子调优、场景适配与融合实现。
- 围绕 Ascend C 的 FlashAttention 算子开展特定场景的性能优化与适配，并研究 PromptFlashAttention 融合算子的优化路径。
- 使用算子测试框架评估待交付算子的精度与性能，并整理评测报告。
- 面向投机推理场景，结合达芬奇架构完成短序列 FP16 Q 与长序列 INT8 KV 的计算优化，并推进至融合算子 PFA 的整合与泛化。

已发表论文

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

- **研究职责：** 负责 FMPQ 细粒度混合精度量化算法的研究与实现。
- 分析 LLM 激活值的离群分布，提出细粒度混合精度量化算法 FMPQ，使 4-bit 激活与 KV Cache 在较小精度损失下可用。
- 面向 W4A4/W4A8 混合精度计算开发 W4Ax 内核，并构建集成该内核与内存管理优化的 COMET 推理框架。
- 在 A100-80G-SXM4 上，相较旧简历所述对比框架实现 2.02× 端到端性能提升；本人负责 FMPQ 的研究与实现。

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

- **研究职责：** 负责结合 MLP 剪枝算法的冷热神经元高效预测。
- 提出将热神经元映射至 GPU、冷神经元卸载至 NDP-DIMM 的异构推理方案，以兼顾计算效率与内存容量。
- 设计实时冷热神经元分区、动态映射与窗口式在线调度机制，实现多 NDP-DIMM 模块间的负载均衡。
- 本人负责结合 MLP 剪枝算法的冷热神经元高效预测。

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

- **研究职责：** 负责硬件友好的动态混合精度量化算法研究，以及动态量化加速器的设计与实现。
- 提出模型无关的低成本动态混合精度量化算法：将 INT8 模型中约 80% 的数据以 INT4 执行，模型精度损失约 1%。
- 设计支持动态精度执行与在线调度的 Drift 加速器；相较 BitFusion，性能提升 2.85×、能耗降低 3.12×。
- 本人负责硬件友好的动态混合精度量化算法研究，以及动态量化加速器的设计与实现。

奖项与荣誉

- 2026 CANN 生态贡献者证书—核心开发者
- 2025 Open Hardware 2025 获奖项目
- 2025 CCF Sys2025 图计算系统设计大赛特等奖（团队成员）