Cai Yiwen

(+86) 189-2876-9793 | Beijing, China | caiyiwen.cs@foxmail.com | github.com/yiwei-cai | yiwen-cai.github.io

Education

Beijing University of Posts and Telecommunications (BUPT) Beijing, China
Master's degree in Computer Science and Technology (Recommended Admission) *Sep 2026 — Present*

Beijing University of Posts and Telecommunications (BUPT) Beijing, China
Bachelor's degree in Computer Science and Technology *Sep 2022 — Jun 2026*

- GPA: 3.76 / 4.0
- Grade: 89.37 / 100.0
- Rank: 59 / 389 (15.16%)
- English: CET6 552 / 710

Work Experience

LLM High-Performance Inference Framework Scheduling & Operator Optimization Research Jul 2026 — Present
Tencent Hunyuan AI Infra *Beijing, China*

- Optimized inference framework scheduling and high-performance operators for LLM serving
- Developed operators including Blackwell-architecture BF16 paged KV attention prefill, nvfp4_blockwise_gemm, per_group MXFP8 quantization, and fused_silu_per_token_quant fusion kernel
- Main open-source repository: Tencent/hpc-ops
- Part of the contributed code has been merged into the vLLM main branch

Projects

The 10th Huawei ICT Competition China Finals Challenge Track Apr 2026 — Apr 2026
Huawei *Huazhong University of Science and Technology, Wuhan*

- Ranked 8th among 64 university teams to advance to finals
- Ranked 3rd overall among 16 finalist teams, awarded First Prize
- Achieved best optimization among all teams in the operator migration track as the sole developer
- Result: National First Prize, Best Operator Migration Optimization Award
- Projects (selected):
  - **fused_add_rmsnorm_kernel**
    - A Triton-based fused RMSNorm and add kernel
    - Multi-row program mapping
    - hidden_size bucketed scheduling
    - 19.28x speedup vs. PyTorch baseline
  - **xllm-ictfinal**
    - Adapted Qwen3.5 model inference on the xllm inference engine
    - Implemented GDN linear attention adaptation
    - Implemented MTP speculative decoding
    - Resolved KVCache memory anomaly
    - Successfully ran 32k-token input inference

cuda-gemm Jun 2025

- Hand-written 8 CUDA GEMM kernels from scratch, covering FP32 CUDA Core and FP16 Tensor Core WMMA
- Progressive optimizations: Shared Memory Tiling, Bank Conflict Avoidance, Double Buffering, cp.async
- SGEMM v3 achieves 95.1% of cuBLAS (4096^2: 37,898 GFLOPS; 8192^2: 36,799 GFLOPS)
- HGEMM v3 Tensor Core reaches 213 TFLOPS at 4096^2, with Nsight Compute/Systems profiling

NUS Summer Workshop Jul 2024 — Aug 2024
National University of Singapore *Singapore*

- Participated in Cloud Computing and Big Data course and completed a group-designed project
- Grade: A+
- Rank: 1st
- Project: **UPLION** Jul 2024 — Aug 2024

- A unified platform for integrating various AI agents for comprehensive AI operations and deployment
- Cloud-native design for scalability and resilience
- Intelligent task distribution and resource optimization
- Kubernetes integration for efficient worker node management

FastGNN May 2026

- A TensorCore-based graph neural network training system
- H100 full-batch GNN
- Custom sparse operator backend
- Voltrix block-sparse SpMM
- Fused3S sparse attention
- Rabbit node reordering
- End-to-end benchmark pipeline
- Validated against DGL/PyG, 1.8x–3.3x speedup

Awards

First Prize, Best Operator Migration Optimization Award, The 10th Huawei ICT Competition China Finals Challenge Track Apr 2026

A+, NUS Summer Workshop Jul 2024

2nd Prize, 15th Blue Bridge Cup National Software and Information Technology Professional Talent Competition Beijing Region Apr 2024

School-level Scholarship (Junior Year) 2025

School-level Scholarship (Sophomore Year) 2024

School-level Scholarship (Freshman Year) 2023

Skills

- **Programming Languages**: Python, C/C++, CUDA, Triton, Rust, Go
- **Research Directions**: LLM training/inference optimization, operator optimization, distributed & cloud computing, high-performance computing