Ziqing Li

PhD Student — Storage Systems & LLM Inference

(+86) 13955537233 @ d202381502@hust.edu.cn

Storage laboratory at Institute of Wuhan National Research Center for Optoelectronics

School of Computer Science and Technology Huazhong University of Science and Technology

Education

2023 - present	*PhD student in Computer Storage System*, HUST
2019 - 2023	*B.S. in Computer Science and Technology*, Beijing Jiaotong University

Publications

- **Ziqing Li**, Jianxi Chen. "OrchKvCache: Orchestrating LLM KV-Cache across GPU–DRAM–SSD with Attention-Aware Hotness Scheduling." *Under review,* **SC 2026**.
- **Ziqing Li**, et al. "SEER: Schedulable KV-Cache Eviction with a Probabilistic Sizing Bound for Real-Time LLM Decoding." *Under review,* **IEEE RTSS 2026**.
- **Ziqing Li**, et al. "HALO: Algebraically Identity-Preserving KV Tiering for Memory-Bounded Long-Context Inference." *Under review,* **EMNLP 2026** (ARR).
- **Ziqing Li**, et al. "Mining Attention Dynamics: Auditing KV-Saliency Prediction in LLMs." *Under review,* **IEEE ICDM 2026**.
- **Ziqing Li**, et al. "Direction Is Not a Knob: A Cross-Generation Cost Characterization of Cross-GPU KV-Cache Handoff Contention on NVLink." *Under review,* **IEEE ICCD 2026**.

Research experience

Long-Context LLM Inference via Tier-Aware KV-Cache Management Apr. 2025 - Present

Advisor: ***Assoc. Prof. Jianxi Chen & Prof. Dan Feng****, School of Computer Science and Technology, HUST*

- **OrchKvCache (SC'26):** attention-driven hot/cold KV tiering over a GPU–DRAM–NVM–SSD hierarchy (atop OrchFS); cuts unnecessary KV migrations 139–597× with 100% bit-exact output, lifts SSD write utilization from 4% to 41% via granularity-aware batching, at P99 scheduling latency $< 60\mu$s
- **SEER (RTSS'26):** recast KV eviction as soft real-time scheduling; a closed-form deadline-miss bound inverts into an SLO-to-KV-budget sizing rule, and a learned attention predictor (P99.9 $\leq 34\mu$s) yields 18/120 SLO-actionable deployment cells vs. 2/120 for raw attention
- **HALO (EMNLP'26):** proved *selection-rule invariance* and built a chunked log-sum-exp KV offload algebraically identical to full attention, enabling 65K-token context on a 24GB GPU (and 32B models on one 80GB GPU) where full attention runs out of memory
- **XQP / KVSalienceBench (ICDM'26):** a 43.9M-example audit of KV-saliency prediction; information-theoretic analysis shows query–key affinity is redundant with attention magnitude, and a 3-parameter calibrated model matches gradient-boosted trees (AUC 0.95, 24× better calibration, 28.7μs P99.9)
- **PeerKV (ICCD'26):** cross-GPU KV-transfer measurement study showing transfer *direction* is a non-effect and victim cost scales with HBM-read footprint (host 1% / peer 9% / local 127%); naive timing manufactures a 74–99pp phantom asymmetry, validated across A100 and H100
- **Ongoing:** query-aware KV-importance prediction, a mixed-tier PagedAttention kernel for vLLM, unified-memory inference on Grace-Hopper / Apple Silicon, and GPU-Direct Storage (GDS) for direct GPU–SSD data paths

Work experience

Big data computing engine development Nov. 2023 - Feb. 2024

Advisor: ***Senior SDE, Apache committer Yaodong Zhang****, Data Center Department, OLAP Engine Group, Xiaomi*

- Implemented Parquet-level encryption and decryption modules within Spark, enabling column-level data protection for internal OLAP pipelines

Awards

- **Huazhong University of Science and Technology First-class academic scholarship** Sep. 2023-25
- **Merit Student Award of Beijing Jiaotong University** Dec. 2020
- **Beijing Jiaotong University Excellent Second Class Scholarship** Dec. 2021

Skills

Programming	C/C++, CUDA, Python, Java, Scala
Systems & Tools	Git, GDB, SPDK, GDS, RDMA, vLLM, PyTorch, HDFS
Research Areas	LLM Inference Systems, KV-Cache Management, Tiered Storage, GPU Systems