

马晓晨

✉ xiaochen.ma.cs@gmail.com · xiaochen.ma@connect.ust.hk
🏠 ma.xiaochen.world · ☎ SunnyHaze · 🎓 Google Scholar · in LinkedIn

研究方向

大语言模型数据基础设施、分布式数据处理系统，以及基于 Ray 和 Apache Spark 的可扩展数据准备。

教育经历

香港科技大学

中国香港

计算机科学与工程博士

2026 年 2 月 – 预计 2028 年

导师：袁彬航教授、张文涛教授。

四川大学

中国成都

计算机科学与技术工学学士

2019 年 9 月 – 2023 年 6 月

本科毕业论文获评四川大学优秀本科毕业论文。

科研与实习经历

腾讯混元基础模型部

中国北京

研究实习生，青云计划

2026 年 6 月 – 至今

- 基于 Ray 和 Apache Spark 构建面向大规模 LLM 数据准备的数据基础设施，重点探索易使用、易分发、易复现的数据算子管理。
- 开发面向数据去重、聚类及大规模文档与视频处理的分布式流水线。

北京大学

中国北京

研究助理，导师：张文涛教授

2025 年 1 月 – 2026 年 1 月

- 负责 DataFlow 项目研发，协调 10 余名同学参与系统设计、开发、代码审查、集成与发布。
- 研发面向大语言模型的数据准备系统，工作涵盖流水线与算子抽象、存储与服务接口，以及基于 Ray 的分布式执行。

四川大学

中国成都

全职研究助理，导师：周吉喆教授

2023 年 7 月 – 2024 年 6 月

- 共同设计并实现 IMDLBenCo，构建模块化的图像篡改检测与定位基准及代码库。
- 研究图像取证、可复现实验评测及高效视觉架构，相关成果发表于 NeurIPS、AAAI 和 ICCV。

代表性系统与开源项目

DataFlow [代码] [论文]

项目负责人

- 面向 LLM 数据准备的统一框架，提供近 200 个可复用算子和 6 条通用流水线，覆盖数据生成、清洗、评估与编排；已投稿 SIGMOD'27。
- 参与核心架构设计与落地并协调 10 余人团队；工程工作涵盖流水线编译与执行、算子注册、存储、服务、CLI 工具及分布式执行。

RayOrch [代码] Flash-MinerU [代码]

发起人及核心开发者

- 构建轻量级异步 Ray 流水线编排抽象，实现微批次重叠执行与 DAG 调度，并将其集成至 DataFlow。
- 为 MinerU 的 PDF-to-Markdown 处理开发可扩展的 Ray 加速层，将串行阶段重构为异步、多 GPU、多节点流水线。

代表性论文

* 共同贡献。完整论文列表见 [Google Scholar](#)。

1. Hao Liang*, **Xiaochen Ma***, et al. *DataFlow: An LLM-Driven Framework for Unified Data Preparation and Workflow Automation in the Era of Data-Centric AI*. 投稿 **SIGMOD'27**。[论文] [代码]
2. **Xiaochen Ma***, Xuekang Zhu*, Lei Su*, Bo Du*, Zhuohang Jiang*, Bingkui Tong*, Zeyu Lei*, Xinyu Yang*, Chi-Man Pun, Jiancheng Lv, and Jizhe Zhou. *IMDL-BenCo: A Comprehensive Benchmark and Codebase for Image Manipulation Detection and Localization*. **NeurIPS'24 Spotlight**。[论文] [代码]
3. Xuekang Zhu*, **Xiaochen Ma***, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Jizhe Zhou. *Mesoscopic Insights: Orchestrating Multi-scale and Hybrid Architecture for Image Manipulation Localization*. **AAAI'25**。[论文] [代码]
4. Bo Du*, Xuekang Zhu*, **Xiaochen Ma***, Chenfan Qu*, Kaiwen Feng*, Zhe Yang, Chi-Man Pun, Jian Liu, and Ji-Zhe Zhou. *ForensicHub: A Unified Benchmark and Codebase for All-Domain Fake Image Detection and Localization*. **NeurIPS'25**。[论文] [代码]
5. Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, **Xiaochen Ma**, et al. *Learning What Reinforcement Learning Can't: Interleaved Online Fine-Tuning for Hardest Questions*. **ICLR'26**。[论文]
6. Yuran Wang, Bohan Zeng, Chengzhuo Tong, Wenxuan Liu, Yang Shi, **Xiaochen Ma**, Hao Liang, Yuanxing Zhang, and Wentao Zhang. *Scone: Bridging Composition and Distinction in Subject-Driven Image Generation via Unified Understanding-Generation Modeling*. **CVPR'26 Highlight**。[论文]
7. Bo Du, **Xiaochen Ma**, Xuekang Zhu, et al. *Can We Build a Monolithic Model for Fake Image Detection? SICA: Semantic-Induced Constrained Adaptation for Unified-Yet-Discriminative Artifact Feature Space Reconstruction*. **ICML'26**。[论文] [代码]
8. Jizhe Zhou, **Xiaochen Ma**, Xia Du, Ahmed Y. Alhammadi, and Wentao Feng. *Pre-training-free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning*. **ICCV'23**。[论文]

学术服务

会议审稿: AAAI'27; NeurIPS'26、BMVC'26、ECCV'26、CVPR'26、ICLR'26、AISTATS'26;
AAAI'25、NeurIPS'25、ICCV'25、ICML'25、CVPR'25、ICLR'25、AISTATS'25; NeurIPS'24、ACM MM'24。

荣誉与奖励

优秀本科毕业论文, 四川大学
全国二等奖, 中国软件杯大学生软件设计大赛

2023 年 6 月
2022 年 10 月

技术能力与语言

- 编程与系统: Python、C/C++、Ray、Apache Spark、PyTorch、Docker、Linux、Git、Bash、 $\LaTeX$ 。
- 语言: 中文 (母语); 英语 (IELTS 7.0, GRE 324)。