

XIAOCHEN MA

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

RESEARCH INTERESTS

LLM data infrastructure, distributed data processing systems, and scalable data preparation with Ray and Apache Spark.

EDUCATION

The Hong Kong University of Science and Technology (HKUST)

Hong Kong SAR, China

Ph.D. in Computer Science and Engineering

Feb. 2026 – Expected 2028

Advisors: [Prof. Binhang Yuan](#) and [Prof. Wentao Zhang](#).

Sichuan University

Chengdu, China

B.Eng. in Computer Science and Technology

Sep. 2019 – Jun. 2023

Outstanding Undergraduate Thesis, Sichuan University.

RESEARCH AND INDUSTRY EXPERIENCE

Tencent Hunyuan, Foundation Model Department

Beijing, China

Research Intern, Qingyun Program

Jun. 2026 – Present

- Building data infrastructure for large-scale LLM data preparation on Ray and Apache Spark, with an emphasis on usable, distributable, and reproducible operator management.
- Developing distributed pipelines for data deduplication, clustering, and large-scale document and video processing.

Peking University

Beijing, China

Research Assistant, advised by [Prof. Wentao Zhang](#)

Jan. 2025 – Jan. 2026

- Led the development of [DataFlow](#), coordinating a team of more than 10 contributors across system design, implementation, code review, integration, and releases.
- Developed data-preparation systems for LLMs, including pipeline and operator abstractions, storage and serving interfaces, and Ray-based distributed execution.

Sichuan University

Chengdu, China

Full-time Research Assistant, advised by [Prof. Jizhe Zhou](#)

Jul. 2023 – Jun. 2024

- Co-led the design and implementation of [IMDLBenCo](#), a modular benchmark and codebase for image manipulation detection and localization.
- Conducted research on image forensics, reproducible evaluation, and efficient vision architectures, resulting in publications at NeurIPS, AAI, and ICCV.

SELECTED SYSTEMS AND OPEN-SOURCE PROJECTS

DataFlow [\[Code\]](#) [\[Paper\]](#)

Project Lead

- A unified framework for LLM data preparation with nearly 200 reusable operators and six domain-general pipelines spanning generation, cleaning, evaluation, and orchestration; submitted to **SIGMOD'27**.
- Helped define and deliver the core architecture while coordinating a 10+ member team; contributed to pipeline compilation and execution, the operator registry, storage, serving, CLI tooling, and distributed execution.

- Built lightweight orchestration abstractions for asynchronous Ray pipelines, including overlapped micro-batch execution and DAG scheduling, and integrated them into DataFlow.
- Developed a scalable Ray-based acceleration layer for MinerU PDF-to-Markdown processing, refactoring sequential stages into an asynchronous multi-GPU and multi-node pipeline.

SELECTED PUBLICATIONS

* *Equal contribution. Full list:* [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*. Submitted to **SIGMOD'27**. [Paper] [Code]
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**. [Paper] [Code]
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**. [Paper] [Code]
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**. [Paper] [Code]
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**. [Paper]
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**. [Paper]
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**. [Paper] [Code]
8. Jizhe Zhou, **Xiaochen Ma**, Xia Du, Ahmed Y. Alhammedi, and Wentao Feng. *Pre-training-free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning*. **ICCV'23**. [Paper]

ACADEMIC SERVICE

Reviewer: 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.

HONORS AND AWARDS

Outstanding Undergraduate Thesis , Sichuan University	<i>Jun. 2023</i>
National Second Prize , China Software Cup College Student Software Design Competition	<i>Oct. 2022</i>

TECHNICAL SKILLS AND LANGUAGES

- **Programming and systems:** Python, C/C++, Ray, Apache Spark, PyTorch, Docker, Linux, Git, Bash, $\text{\LaTeX}$.
- **Languages:** Mandarin (native); English (IELTS 7.0, GRE 324).