Guorui Xiao

Mobile: 206-502-9268

Email: grxiao@cs.washington.edu
Website: xertxiao.github.io

Research Interest

I am interested in data management, machine learning systems, large language models, and diffusion models.

Education

- **University of Washington** — Seattle, WA, USA
  Ph.D. Student - Computer Science; GPA: 3.93/4.0 — *Expected Graduation:: Sep. 2028*
  Advisor: Magdalena Balazinska
- **University of California, Los Angeles** — Los Angeles, CA, USA
  Masters of Science - Computer Science; GPA: 4.0/4.0 — *Graduated:: Mar. 2023*
  Advisor: Carlo Zaniolo
- **University of California, Los Angeles** — Los Angeles, CA, USA
  Bachelor of Science - Computer Science; GPA: 3.77/4.0; Cum Laude — *Graduated: Dec. 2020*

Selected Publications & Manuscripts

[P1] **CENTS: A Flexible and Cost-Effective Framework for LLM-Based Table Understanding**
Guorui Xiao, Dong He, Jin Wang, Magdalena Balazinska. International Conference on Very Large Data Bases (**VLDB**) 2025, pages 4574-4587.

[P2] **Highly Efficient String Similarity Search and Join over Compressed Indexes**
Guorui Xiao, Jin Wang, Chunbin Lin, Carlo Zaniolo. IEEE International Conference on Data Engineering (**ICDE**) 2022, pages: 232-244.

[P3] **RACOON: An LLM-based Framework for Retrieval-Augmented Column Type Annotation with a Knowledge Graph**
Linxi Wei, **Guorui Xiao**, and Magdalena Balazinska. Neural Information Processing Systems (**NeurIPS**) 2024, Table Representation Learning Workshop.

[P4] **Revealing Protocol Architecture's Design Patterns in the Volumetric DDoS Defense Design Space**
Zhiyi Zhang, **Guorui Xiao**, Sichen Song, Angelos Stavrou, Eric Osterweil, and Lixia Zhang. IEEE Communications Surveys and Tutorials 2024. (survey paper)

[P5] **RaSQL: A Powerful Language and its System for Big Data Applications**
Jin Wang, **Guorui Xiao**, Jiaqi Gu, Jiacheng Wu, Carlo Zaniolo. ACM International Conference on Management of Data (**SIGMOD**) 2020, pages: 2673-2676. (demo paper)

[M1] **KathDB: Explainable Multimodal Database Management System with Human-AI Collaboration**
Guorui Xiao, Enhao Zhang, Nicole Sullivan, Will Hansen, and Magdalena Balazinska. Under review by Conference on Innovative Data Systems Research (**CIDR**) 2026.

[M2] **RACOON^{+}: A System for LLM-based Table Understanding with a Knowledge Graph**
Linxi Wei, **Guorui Xiao**, Moe Kayali, Dan Suciu, and Magdalena Balazinska. Under review by International Conference on Very Large Data Bases (**VLDB**) 2026.

[M3] **RS-SQL: A Query Language for Supporting Recursive Query Processing over Data Streams**
Guorui Xiao, Jin Wang, Jiacheng Wu, Carlo Zaniolo. To be submitted to The International Journal on Very Large Data Bases (**VLDB Journal**)

Ongoing Research Projects

- **UW Database Group (UWDB)** — University of Washington
  Research Assistant — *Sep. 2025 - Present*
  - **Agentic Multi-Modal Database Management System [M1]**
    - Designed and implemented a prototype DBMS powered by LLM agents that answers natural language queries while providing robust, traceable, and explainable results.
    - Introduced a semantic middle layer that defines modality-aware views, enabling uniform query semantics across heterogeneous data sources.
    - Proposed an LLM-generated Function-as-Operator (FAO) abstraction that generates implementations of physical operators on-the-fly, improving operator-level explainability and integrating provenance tracking during query execution.
    - Demonstrated through preliminary experiments how tuple-level lineage can be reconstructed via physical plan and intermediate result materialization.

Selected Previous Research Projects

- **UW Database Group (UWDB)** — University of Washington
  Research Assistant — *Dec. 2023 - Jun. 2025*
  - **Large Language Model for Table Understanding**
    - **Unified cost-effective LLM-based framework for Table Understanding [P1]**
      - Proposed a unified, cost-effective LLM-based framework for table understanding, optimizing performance while managing token costs within budget constraints.
      - Devised two novel yet general reducer paradigms to reduce table content and task-specific labels, minimizing misleading information and enhancing LLM comprehension.
      - Achieved performance improvements of up to 0.21 micro F1 in column type annotation, 0.12 micro F1 in relation extraction, and 0.29 mean Average Precision in schema augmentation compared to baseline models.
    - **Using RAG to improve LLM performance on Column Type Annotation [P2, M2]**
      - Led a research intern to develop a novel system RACOON that augments prompts for LLMs using external non-parametric knowledge from a knowledge graph through RAG for a variety of table understanding tasks, achieving up to 0.21 improvement in micro F1 against baselines.

- **Scalable Analytics Institute (ScAi)** — University of California, Los Angeles
  Research Intern — *Dec. 2019 - Mar. 2023*
  - **Streaming Data Processing System that Supports Recursive Queries [M3]**
    - Proposed a high-level query language by drawing inspirations from Datalog for data streams to support expressing recursive queries.
    - Devised a lightweight structure *Queue-Based Index* to avoid redundant computation and further proposed an efficient query evaluation method based on it.
    - Designed and implemented a prototype datastream system to verify the effectiveness of the designs.
    - Conducted experiments that showed improvements of ∼10X in throughput and ∼5X in tail latency.
  - **Unified Compression Framework to Support String Similarity Queries [P2]**
    - Proposed the first unified framework for offline and online construction of compressed inverted index to support String Similarity Search/Join applications to avoid expensive disk I/O costs.
    - Devised algorithms to achieve near-optimal compression ratio in an online manner with tools like Kernel Density Estimation.
    - Conducted experiments that showed we improved ∼5X in memory consumption.

- **Internet Research Laboratory (IRL)** — University of California, Los Angeles
  Research Intern — *Jun. 2020 - Sep. 2020*
  - **Systematization of Knowledge: distributed denial-of-service (DDoS) attack [P4]**
    - Systematically selected ∼250 papers out of ∼24,000 works related to volumetric DDoS attack and closely examined ∼50 of them.
    - Performed detailed analysis over selected to derive systematized repeating design patterns and a set of IP network architecture properties.
    - Categorized the above papers into sub-categories based on their deployment locations, approaches, incentives, etc. and contributed to writing a research paper.

Selected Industry Experience

- **Apple, Inc.** — Seattle, WA, USA
  Research Intern — *Jun. 2025 - Sep. 2025*

  Multi-Modal and Multi-Task Vision Foundation Model Joint Training
  - Designed and implemented a joint training framework that unifies image and video data for a DiT-based video foundation model trained with rectified flow, using flexible data recipes.
  - Enabled a single foundational model to simultaneously support text-to-image, text-to-video, and image-to-video generation.
  - Conducted experiments demonstrating that this paradigm not only enable new capabilities but also improved spatial metrics (e.g., Imaging Quality) and temporal metrics (e.g., Dynamic Degree) compared to a baseline model trained on a single modality and single task.

- **Arista Networks, Inc.** — Los Angeles, CA, USA
  Software Engineer Intern — *Jun. 2022 - Sep. 2022*

  IEEE 802.1Q Tunneling CLI
  - Designed the new module architecture that significantly reduced the code complexity compared to the existing similar tunneling implementation and completed a detailed design document.
  - Implemented software-side reactors and hardware-side bit setter that together can filter packets violating user-defined VLAN rules in 802.1Q tunneling.

- Pushed the changes to the next release to be used by all switches over a specific popular platform.

- **Taboola, Inc.** Los Angeles, CA, USA
  Machine Learning/Data Science Intern *Jun. 2019 - Sep. 2019*
  Knowledge Base of News Keywords
  - Built an end-to-end pipeline with Spark SQL and Java to process data crawled by IBM Watson and further capture their embeddings with Word2Vec.
  - Devised algorithms for de-duplicating keywords based on a combined metric, including similar neighbors, lexical similarity, etc.
  - Proposed a Knowledge Base representation of news keywords over Neo4j to effectively visualize keywords relationships and implemented an auto-renewal process that runs daily.

Teaching Experience

- **COM SCI 35L: Software Construction Laboratory** Los Angeles, CA, USA
  Teaching Assistant *Fall 2021*
  - Lectured 20 hours of material focusing on Git, Shell, Vim, Java, etc., to 52 students and held 20 hours of office hours for ∼250 students.
  - Mentored ∼10 groups of undergraduate students completing Node.js/React projects.
  - Graded ∼250 students' coding assignments and 2 exams.

Mentoring Experience

- **Undergraduate Research Mentorship** Los Angeles, CA, USA
  Lindsey Linxi Wei *Jun. 2024 - Jun. 2025*
  - Mentored an undergraduate student, guiding her research project that produced two first-author papers: one published at NeurIPS 2024 [P3], and another under review at VLDB 2026 [M2].

Misc

- **Selected Languages**: Python, C/C++, Java, SQL, Bash, Datalog.
- **Selected Platforms**: Amazon EC2, Sklearn, Github, Neo4j, Apache Spark, Apache Flink, Spark Streaming, LangGraph, LaTeX.