# Xinyue Ma

Email: xinyuema@postech.ac.kr PhD Student

## RESEARCH INTERESTS

- Systems for efficient and scalable large language model inference
- Resource-efficient continual and on-device learning

## EDUCATION

**Pohang University of Science and Technology (POSTECH)** Aug. 2024 – Present

Combined Master and PhD in Artificial Intelligence
Affiliation: OMNIA Lab of POSTECH
Advisor: Myeongjae Jeon

**Ulsan National Institute of Science and Technology (UNIST)** Feb. 2022 – Aug. 2024

Combined Master and PhD in Computer Science Engineering
Affiliation: OMNIA Lab of UNIST
Advisor: Myeongjae Jeon

**Korea Advanced Institute of Science and Technology** Feb. 2017 – Aug. 2021

B.S. in Electrical Engineering | Minor: Computer Science
Advisor: Euijong Steven Whang

## PUBLICATIONS

1. ORBITFLOW: SLO-Aware Long-Context LLM Serving with Fine-Grained KV Cache Reconfiguration — *VLDB* 2026 under review
**Xinyue Ma***, Heelim Hong*, Jongseob Lee, Seoyeong Choy, Woo-Yeon Lee, Taegeon Um, Myeongjae Jeon

2. REP: Resource-Efficient Prompting for Rehearsal-Free Continual Learning — *NeurIPS* 2025
Sungho Jeon, **Xinyue Ma**, Kwang In Kim, Myeongjae Jeon

3. Cost-effective On-device Continual Learning over Memory Hierarchy with Miro — *ACM MobiCom* Oct. 2023
**Xinyue Ma**, Suyeon Jeong, Minjia Zhang, Di Wang, Jonghyun Choi, Myeongjae Jeon

## RESEARCH PROJECTS

**Fast and Efficient Long-Context LLM Serving**
* Extension of an internship project at Microsoft Research

Retrieval-augmented generation (RAG) and shared knowledge bases are widely adopted to reduce hallucinations and improve generation quality, but they introduce new system challenges: prompts become much longer after injecting retrieved contexts, and LLMs repeatedly compute attention over the same shared knowledge across users and requests. LLM inference naturally splits into a compute-bound prefill stage, where input tokens are processed in parallel to build context, and a memory-bound decode stage, where frequent KV-cache accesses grow with context length and the number of generated tokens. My recent work, OrbitFlow, targets the decode stage in memory-constrained settings, providing adaptive, dynamic KV cache management over multi-tier storage so that KV placement across tiers is continuously adjusted to satisfy latency SLOs under tight memory limits. My current work focuses on the compute-bound prefill stage, aiming to make it faster through efficient KV reuse beyond simple prefix sharing and selective recomputation of only the necessary parts of the KV cache.

2024.09 – Present

**Automatic 3D-Parallelization for Deep Learning with Heterogeneous GPUs**
* Collaboration with Samsung Research

As GPUs become more powerful and diverse, training large-scale models on clusters of heterogeneous GPUs is increasingly unavoidable. However, existing parallelization schemes (data, tensor, pipeline parallelism) largely assume homogeneous hardware and fail to fully exploit the heterogeneity in real clusters, leaving substantial potential efficiency on the table. In this project, we aim to develop a framework that automatically explores 3D-parallelization strategies tailored to the actual mix of GPUs in the cluster. By modeling the cost of different combinations of pipeline, tensor, and data parallelism and searching this space efficiently, the framework selects cost-effective strategies that minimize training time and resource usage for large LLM workloads.

2023.06 – 2024.02

**Cost-effective On-device Continual Learning Systems**
* Collaboration with DeepSpeed and Yonsei University

DL applications on the edge often require training DNN models (1) within the edge device to preserve privacy and (2) continually as data arrives, an emerging learning paradigm for edge devices known as continual learning. With limited resources on edge devices, cost-efficiency becomes critical for such training jobs, but prior studies are predominantly simulation-based and give limited attention to cost-efficiency. Together with my colleagues, I focus on system solutions for cost-effective on-device continual learning, achieving high model accuracy without compromising energy efficiency. Our work Miro dynamically optimizes CL configurations (e.g., memory and I/O usage) based on resource states to exploit favorable accuracy–energy tradeoffs. Building on this, REP improves the computational and memory efficiency of transformer-based CL models so that they become practical on resource-constrained edge devices.

2023.04 – 2024.09

## EXPERIENCES

**Intern**, Research in Software Engineering Group
Microsoft Research Redmond, U.S.A — Sep. 2024 – Nov. 2024

**Intern**, Intelligent Edge and Cloud Group
Microsoft Research Asia, China — Nov. 2023 – Feb. 2024

**Teaching Assistant**, POSTECH, Korea
AI Systems — Fall 2024

**Teaching Assistant**, UNIST, Korea
Computer Architecture — Fall 2023
Parallel Computing — Fall 2022
Advanced Programming — Spring 2022