Yixin Dong

yixind@andrew.cmu.edu · Ubospica · @yi_xin_dong

Education

Carnegie Mellon University, Pittsburgh, USA 2024.9 – Present

Ph.D. in Machine Learning (Advisor: Prof. Tianqi Chen)

Shanghai Jiao Tong University, Shanghai, China 2020.9 – 2024.6

B.S. in Computer Science (ACM Honors Class, an honor program for top 5% students), Zhiyuan College

- GPA: 4.03/4.3 (2/37); *summa cum laude*

Research Interests

Large Language Models, LLM Agents, Machine Learning Systems, Machine Learning Compilers

Open Source Projects

Apache TVM — Github ★ 12.2k 3.6k 2022.8 – Present

- Open-source machine learning compiler enabling high-performance model deployment across heterogeneous hardware backends
- Lead developer of TVM Unity-AD, the next-generation automatic differentiation framework featuring compilation optimizations and cross-platform deployment
- Contributed major components, including TVM Unity IR, DLight GPU Scheduler, and TVM Runtime
- Serve as an official Apache TVM Reviewer, maintaining core modules and driving design discussions

MLC-LLM — Github ★ 20.4k 1.7k 2023.2 – Present

- Compile and deploy LLMs natively and efficiently on diverse platforms, including laptops, Macs, iPhones, and Android devices
- Major contributor responsible for GPU kernel, API integration, and LLM code generation

XGrammar — Github ★ 1.4k 101 **29M** 2024.11 – Present

- Structured generation engine producing outputs in user-specified formats with 100% format correctness, enabling reliable downstream applications
- Achieves more than 10× speedup over previous state-of-the-art structured generation solutions
- Core component of major LLM serving frameworks: SGLang, MLC-LLM, vLLM, TensorRT-LLM, etc.
- Widely adopted by industry partners: Google, Meta, Alibaba, xAI, DeepSeek AI, Databricks, Perplexity AI, Modular AI, etc.

FlashInfer-Bench — Github 2025.10 – Present

- First closed-loop framework for rigorous evaluation and deployment of AI-agent-generated GPU kernels
- Standardizing kernel-generation protocol, enforcing strict correctness/performance evaluation, and closing the loop of kernel deployment
- Project lead; supervised a team of three undergraduate researchers
- Adopted by major LLM frameworks (SGLang, vLLM, FlashInfer, etc.) and major industry partners (NVIDIA, Bosch, etc.)

Publications

XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models
MLSys 2025

***Yixin Dong**, Charlie F Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, Tianqi Chen*
Paper | Github

Experiences

xAI 2025.8 – Present

Part-time Researcher

- **Tool Calling and LLM Agents**
  - Tool calling and structured generation support for Grok-4 and Grok-Code-Fast-1

Databricks 2025.5 – 2025.8

Research Intern

- **Enterprise-oriented Agent Workflow**
  - Built the structured workflow for Agent Bricks, Databricks' major enterprise agent platform
  - Built an LLM speculative decoding system with LoRA Customization

DeepSeek 2024.3 – 2024.7

Research Intern

- **LLM Pretraining and Inference Optimization**
  - Involved in the training and inference optimization of Deepseek-V2, a strong Mixture-of-Experts (MoE) language model characterized by economical training and efficient inference
  - Designed a code generation engine for Deepseek-V2-Coder, enhancing code generation and function-calling capabilities

SAMPL, University of Washington 2023.7 – 2024.1

Research Intern, advised by Prof. Luis Ceze and Prof. Tianqi Chen (CMU)

- **On-device Deployment of Large Language Model Fine-tuning**
  - Enabled fine-tuning of large language models on Mac, AMD, and iPhone GPUs (for the first time)
- **Efficient GPU Kernel Generation**
  - Designed an automatic GPU kernel generation framework for LLMs, achieving state-of-the-art performance on both NVIDIA and Apple GPUs

Apex Lab, Shanghai Jiao Tong University 2022.8 – 2023.6

Research Intern, advised by Prof. Yong Yu and Prof. Weinan Zhang

- **Training Optimization for Machine Learning Compilers**
  - Designed TVM Unity-AD, a next-generation automatic differentiation framework for Apache TVM

Honors

Programming Competitions

Gold Medal, 2020 ICPC Asia East Continent Final 2021.4

Gold Medal, 2020 ICPC Asia Shanghai Regional Contest 2020.11

Gold Award, 2020 China Collegiate Programming Contest, Mianyang Site 2020.11

Scholarships

Fan Hsu-Chi Chancellor's Scholarship 2023.8, 2022.8

One of the highest-level scholarships at SJTU, awarded to the top 0.1% students (15 out of 15,000)

China National Scholarship 2021.11

Awarded to top 0.2% students nationwide

Hanyingjuhua Outstanding Student Scholarship 2021.11
Zhiyuan Honor Scholarship 2022.12, 2021.12, 2020.12

Selected Individual Projects

CompilerStorm 2022.1

A compiler designed from scratch that compiles a C++-like language into RISC-V assembly, featuring JIT compilation and register allocation optimization

Hummingbird 2021.11

A RISC-V processor implemented in Verilog with out-of-order execution and branch prediction, running on an FPGA development board

Selected Talks

FlashInfer-Bench: Building the Virtuous Cycle for AI-driven LLM Systems, CMU, NVIDIA, Databricks, Bosch 2025

XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models, CMU; THU; UC Berkeley; MIT; Rice; LMSys; Ant Group; UC Davis 2024, 2025

Universal Deployment of LLM Finetuning, UW 2023.11

On-device Training on Machine Learning Compiler, SJTU 2023.4

Cross-platform Training Using Automatic Differentiation on Relax IR, TVMCON 2023.3

Leadership

Co-organizer of the 2023 ACM-Class Student Academic Festival (ASAF2023) 2023.6

Coordinated the conference schedule and helped invite seven professors and 14 Ph.D. students worldwide

Team Leader of the Overidea team in ACM-ICPC 2020.9-2021.6

Teaching

Operating Systems, taught by Prof. Alei Liang 2023.3-2023.7

Teaching assistant; led the design of ACMOS, an education-oriented OS

C++ Programming, taught by Prof. Huiyu Weng 2021.9-2022.1

Teaching assistant (leading)