

# YIXIN DONG

✉ [yixind@andrew.cmu.edu](mailto: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 deployment models on diverse hardware backends
- Leading the TVM Unity-AD project, the next-generation automatic differentiation framework featuring compilation optimization and cross-platform deployment
- Contributing to several key features: TVM Unity IR, DLight GPU Scheduler, TVM Runtime
- Serving as Apache TVM Reviewer in the community

**MLC-LLM** —  [Github](#) ★ 20.4k  1.7k 2023.2 – Present

- Compile and deploy LLMs natively and fast on various platforms, including laptops, Macs, iPhones, and Android devices
- Main contributor to the project, contributing to GPU kernel generation, API integration, and LLM code generation

**XGrammar** —  [Github](#) ★ 1.3k  86  19M 2024.11 – Present

- An LLM structured generation engine that accepts user-specified format and outputs in this format with 100% accuracy, benefiting downstream applications
- 10x faster than the previous state-of-the-art structured generation solution
- Integrated into major LLM serving engines, including SGLang, MLC-LLM, vLLM, TensorRT-LLM, etc.
- Adopted by xAI, DeepSeek AI, Alibaba (Qwen), Databricks, Perplexity AI, Modular AI, 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

*Research Intern*

- **Tool Calling and LLM Agents**
  - Enhance the tool calling ability 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

---

<b>XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models</b> , CMU; THU; UC Berkeley; MIT; Rice; LMSys; Ant Group	2024, 2025
<b>Universal Deployment of LLM Finetuning</b> , UW	2023.11
<b>On-device Training on Machine Learning Compiler</b> , SJTU	2023.4
<b>Cross-platform Training Using Automatic Differentiation on Relax IR</b> , TVMCON	2023.3

## LEADERSHIP

---

<b>Co-organizer</b> of the 2023 ACM-Class Student Academic Festival (ASAF2023) Coordinated the conference schedule and helped invite seven professors and 14 Ph.D. students worldwide	2023.6
<b>Team Leader</b> of the Overridea team in ACM-ICPC	2020.9-2021.6

## TEACHING

---

<b>Operating Systems</b> , taught by Prof. Alei Liang Teaching assistant; led the design of ACMOS, an education-oriented OS	2023.3-2023.7
<b>C++ Programming</b> , taught by Prof. Huiyu Weng Teaching assistant (leading)	2021.9-2022.1