

# TIAN-YU GUO 郭天宇

✉ guoty9@mail2.sysu.edu.cn · 🔗 “gty111.github.io/info” · 🌐 “github.com/gty111”

## 🎓 EDUCATION

**Sun Yat-sen University**, Guangzhou, China

2022 – 2027 (expected)

*PH.D.* in Computer Science (CS)

**Xidian University**, Shaanxi, China

2018 – 2022

*B.S.* in Computer Science (CS)

## 📖 PUBLICATIONS

- **arXiv’25**: **Tianyu Guo**, Xianwei Zhang, Jiangsu Du, Zhiguang Chen, Nong Xiao, Yutong Lu, “gLLM: Global Balanced Pipeline Parallelism System for Distributed LLM Serving with Token Throttling”
- **Euro-Par’25**: **Tianyu Guo**, Hande Dong, Yichong Leng, Feng Liu, Cheater Lin, Nong Xiao and Xianwei Zhang, “EFIM: Efficient Serving of LLMs for Infilling Tasks with Improved KV Cache Reuse”, The 31st International European Conference on Parallel and Distributed Processing, Dresden, Germany, August.
- **ASP-DAC’25**: Mengyue Xi, **Tianyu Guo**, Xuanteng Huang, Zejia Lin, Xianwei Zhang, “Mpache: Interaction Aware Multi-level Cache Bypassing on GPUs”, The 30th Asia and South Pacific Design Automation Conference, Tokyo Odaiba Miraikan, Japan, January.
- **DAC’24**: **Tianyu Guo**, Xuanteng Huang, Kan Wu, Xianwei Zhang and Nong Xiao, “SMILE: LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism”, The 61st ACM/IEEE Design Automation Conference, San Francisco, CA, United States, June.

## 👥 EXPERIENCE AND PROJECTS

### **gLLM: Global Balanced Pipeline Parallelism System with Token Throttling**

2025

Pipeline parallelism in LLMs often suffers from inefficiencies due to load imbalance, particularly inter-batch imbalance, which existing methods frequently overlook. To address this, we propose Token Throttling, a dynamic strategy that adapts the number of prefill and decode tokens per batch based on real-time system conditions. By reducing computational imbalance between batches, our approach significantly improves GPU utilization. Experiments show that gLLM outperforms vLLM and SGLang, achieving 11% to 398% higher throughput while maintaining lower latency.

### **Efficient Serving of Code Completion LLMs (Tencent Intern)**

2024

Code completion LLMs tend to own the ability of fill-in-the-middle (FIM). We propose EFIM, a transformed prompt format of FIM to unleash the performance potential of KV cache reuse. We also propose an enhanced training procedure on data processing for solving subtoken generation problems. EFIM can lower the latency by 52% and improve the throughput by 98% while maintaining the original code completion capability.

### **Cross-request KV Cache Management (Tencent Intern)**

2024

The LLM inference is used interactively in a multi-round fashion with repeated context information, thereby incurring redundant computation and further prolonged inference stages. We propose KV sail, a cross-request KV cache management to maintain a per-user session to reuse the data in multi-round interactions. KV sail effectively outperforms the state-of-the-art by 37%/190% on throughput and 24%/68% on latency.

### **Fine-grained Kernel Scheduling and Management to Improve GPU Sharing**

2023

Existing GPU sharing adopts either coarse-grained collocation strategies or interference-unaware spatial partition strategies. We propose FEDCM, a kernel-level collocation-based GPU sharing scheme to establish a federated use of compute and on-chip memory resources. FEDCM improves the overall throughput by 48.3% and 17.4%, compared to standard sharing baseline and prior state-of-the-art, respectively.

## **Optimize GEMM Step by Step** 2023

“GEMM MMA” first implementates a naive kernel of GEMM by CUDA `mma.sync` and then optimize it step by step (using vectorization, asynchronous copy, conflict-free shared memory access, consolidated memory access and so on), which achieves above 70% of peak performance relative to CUTLASS in the final version.

## **Teaching Assistant of “SYSU-DCS3013 : Computer Architecture”** 2022

Release “SYSU-ARCH LAB” which focuses on simulators (gem5, GPGPU-Sim and Accel-Sim).

## **Design PTX-EMU** 2022

“PTX-EMU” is an simulator for NVIDIA’s virtual instruction set PTX. You can use it to generate image by simulating rendering program.

## SKILLS

---

- Programming Languages: C, C++, CUDA, Python, Java
- English: CET6 (517/750)

## HONORS AND AWARDS

---

The Second Prize in ACTIC of A3 track operator implementation and performance optimization 2023

Top 4.2% nationwide | Top 2.9% worldwide in “leetcode contest” 2023

The Second Prize Scholarship in SYSU 2022-2023

The Second Prize Scholarship in XDU 2018-2019, 2019-2020, 2020-2021