Shuoming Zhang（张朔铭）

Ph.D. Candidate in Computer Architecture

State Key Laboratory of Processors, Institute of Computing Technology, CAS | University of Chinese Academy of Sciences

zhangshuoming17@mails.ucas.ac.cn • Beijing, China • zhangshuoming990105.github.io

Google Scholar • ORCID 0009-0004-4210-5123 • GitHub • OpenReview

Research Interests

Kernel agents (LLM-driven generation, optimization, and verification of low-level compute kernels); large-scale multi-agent systems; LLM infrastructure and serving systems, with an emphasis on constrained and structured decoding; reliability and safety for LLM-driven systems; compiler design and code translation.

Education

Ph.D. in Computer Architecture, Institute of Computing Technology, CAS & UCAS 2021 – present
Advisors: Prof. Huimin Cui and Assoc. Prof. Jiacheng Zhao. State Key Laboratory of Processors (SKLP). Beijing

B.Eng. in Computer Science, University of Chinese Academy of Sciences (UCAS) 2017 – 2021

Selected Honors & Awards

- **Distinguished Paper Award**, CGO 2026 — *From Threads to Tiles (T2T)*.
- **Spotlight**, ICML 2026 — *CONTINUUM*.
- **Hygon Ph.D. Award**, 2024.

Publications

† equal contribution * corresponding author. Author name in **bold**.

Peer-Reviewed

1. **[ICML 2026, spotlight]** *CONTINUUM: Restoring the Contiguous Tensor Abstraction Efficiently for Dynamic AI Workloads via Hardware Virtualization*
   Yangyu Zhang†, **Shuoming Zhang**†, Chunwei Xia, Shuaijiang Li, Zhicheng Li, Ruiyuan Xu, Zheming Yang, Lei Chen, Yuan Wen, Guangli Li, Xiaobing Feng, Huimin Cui, Jiacheng Zhao*

2. **[ICML 2026]** *LEGO: An LLM-Enabled Hierarchical Optimizer for Tensor Computation Graphs with Structure-Aware Search and Compositional Synthesis*
   Ruiyuan Xu†, **Shuoming Zhang**†, Guangli Li, Qiuchu Yu, Rui Zhang, Yangyu Zhang, Hao Qian, Chunwei Xia, Jiacheng Zhao, Chenxi Wang, Xiaobing Feng, Jingling Xue, Huimin Cui

3. **[CCS 2026]** *When Grammar Guides the Attack: Uncovering Control-Plane Vulnerabilities in LLMs with Structured Output*
   Shuoming Zhang, Jiacheng Zhao*, Hanyuan Dong, Ruiyuan Xu, Zhicheng Li, Yangyu Zhang, Shuaijiang Li, Yuan Wen, Chunwei Xia, Zheng Wang, Xiaobing Feng, Huimin Cui
   arXiv:2503.24191

4. **[ISCA 2026]** *Symbiotic MLLM Serving: Dynamically Balancing Parallelism Across GPUs and Resources Within GPUs*
   Zhicheng Li, Jiacheng Zhao*, Yangyu Zhang, Zhaolin Duan, Xinyu Liu, Siqi Li, **Shuoming Zhang**, Shuaijiang Li, Donglin Yu, Yuan Wen, Chunwei Xia, Xiyu Shi, Huimin Cui

5. **[CGO 2026, Distinguished Paper Award]** *From Threads to Tiles: T2T, a Compiler for CUDA-to-NPU Translation via 2D Vectorization*
   Shuaijiang Li, Jiacheng Zhao*, Ying Liu, **Shuoming Zhang**, Lei Chen, Yijin Li, Yangyu Zhang, Zhicheng Li, Runyu Zhou, Xiyu Shi, Chunwei Xia, Yuan Wen, Xiaobing Feng, Huimin Cui

6. **[SCIS 2026]** *Large Processor Chip Model*
   Kaiyan Chang, Mingzhi Chen, Yunji Chen*, …, Rui Zhang, **Shuoming Zhang**, Jiacheng Zhao (alphabetical order, equal contributions)
   Springer

7. **[CCF THPC]** *The New Compiler Stack: A Survey on the Synergy of LLMs and Compilers*
   Shuoming Zhang, Jiacheng Zhao*, Qiuchu Yu, Chunwei Xia, Zheng Wang, Xiaobing Feng, Huimin Cui
8. **[CCF THPC]** *LEGO-Compiler: Enhancing Neural Compilation Through Translation Composability*
   Shuoming Zhang, Jiacheng Zhao, Chunwei Xia, Zheng Wang, Yunji Chen, Xiaobing Feng, Huimin Cui*
9. **[NeurIPS 2025, poster]** *SpaceServe: Spatial Multiplexing of Complementary Encoders and Decoders for Multimodal LLMs*
   Zhicheng Li, **Shuoming Zhang**, Jiacheng Zhao*, Siqi Li, Xiyu Shi, Yangyu Zhang, Shuaijiang Li, Donglin Yu, Zheming Yang, Yuan Wen, Huimin Cui
   OpenReview
10. **[NeurIPS 2025, poster]** *Mutual-Supervised Learning for Sequential-to-Parallel Code Translation*
    Changxin Ke, Rui Zhang*, Shuo Wang, Li Ding, Guangli Li, Yuanbo Wen, **Shuoming Zhang**, Ruiyuan Xu, Jin Qin, Jiaming Guo, Chenxi Wang, Ling Li, Qi Guo, Yunji Chen
    OpenReview
11. **[EMNLP 2024, findings]** *Introducing Compiler Semantics into Large Language Models as Programming Language Translators: A Case Study of C to x86 Assembly*
    Shuoming Zhang, Jiacheng Zhao, Chunwei Xia, Zheng Wang, Yunji Chen, Huimin Cui*

Preprints

1. **[arXiv 2026]** *Learning When to Optimize: Verified Optimization Skills from Expert GPU-Kernel Lineages (KLineage)*
   Shuoming Zhang†, Qiuchu Yu†, Yangyu Zhang, Ruiyuan Xu, Xiyu Shi, Guangli Li, Xiaobing Feng, Huimin Cui, Jiacheng Zhao*
   arXiv:2605.28213

Research Experience

Current projects

- **Kernel agents.** Agentic pipelines that synthesize, tune, and verify high-performance kernels for AI accelerators with LLM-in-the-loop search and feedback.
- **Large-scale multi-agent systems.** Orchestration and runtime support for populations of cooperating LLM agents on complex software-engineering and system-level tasks.
- **Constrained and structured decoding.** Decoding as LLM infrastructure: constrained and grammar-based decoding for controllable, reliable, and safe generation (e.g., structured-output control-plane analysis).
- **Grammar-guided code generation** *(work in progress).* Steering LLMs with grammar constraints to produce syntactically and semantically valid code.

Earlier projects

- **LLM-guided compilation workflows.** Model-in-the-loop pipelines for source-to-assembly translation and error recovery with LLM feedback.
- **LLM-aware compiler construction.** Reusable compiler components leveraging LLM reasoning for IR transformation, code generation, and verification.
- **Heterogeneous model offloading with TVM** (collaboration with Intel). NPU/CPU co-execution and scheduling within the TVM stack; prototyped a new TVM backend for a simulator-based NPU.
- **VLIW instruction scheduling** (collaboration with Huawei). Instruction-scheduling heuristics targeting domain-specific VLIW architectures.

Professional Service

Reviewer: Transactions on Machine Learning Research (TMLR), Artificial Intelligence Review, ICML, NeurIPS, ICLR.

Technical Skills

Languages: C/C++, Python, CUDA. **Systems & Tools:** LLVM/MLIR, TVM, PyTorch, LLM serving/inference stacks, Git, Linux.

Areas: compiler construction, code generation & translation, GPU/NPU kernel optimization, LLM agents and serving systems.