

Jiayi Tian

+1 (805) 245 0298 | jiayi_tian@ucsb.edu | linkedin.com/in/jiayi-tian-32b9652a5/

Focus on Efficient LLM Training & Inference, Efficient CoT Reasoning via Algorithm-Hardware Co-Design.

EDUCATION

University of California, Santa Barbara, Ph.D. in Computer Engineering | CA, USA **3.9/4.0**

Fall 2025 - ongoing

University of California, Santa Barbara, M.S. in Computer Engineering | CA, USA **3.9/4.0**

Fall 2023 - Fall 2025

Nanjing University, B.Eng. in Electrical Engineering | China **4.5/5.0**

Fall 2019 - Fall 2023

INDUSTRIAL EXPERIENCE

Intel Corporation, Research Intern 2025 | Hillsboro, OR

Jun 2025 – Sep 2025

- Proposed SkipKV, a training-free KV-cache compression framework featuring sentence-level selective eviction and dynamic generation control for efficient chain-of-thought (CoT) reasoning in multi-batch settings.
- Designed a semantic similarity metric to evict redundant sentences from the cache, improving reasoning coherence.
- Introduced a dynamic activation-steering mechanism to enable concise and stable inference.
- Demonstrated strong results on long-reasoning tasks (e.g. AIME24, LiveCodeBench) with LRMs: up to 26.7% higher accuracy, 1.6 \times shorter generation, and 1.7 \times higher throughput vs. SOTA under equal compression.
- Resulting paper submitted to MLSYS.

Intel Corporation, Research Intern 2024 | Hillsboro, OR

Jun 2024 - Sep 2024

- Proposed a tensor-compressed Transformer training accelerator on FPGA, optimizing compute ordering, dataflow, and memory allocation for LLMs.
- Designed a bidirectional tensor contraction scheme enabling substantial reduction in intermediate memory and compute cost during long-sequence training and inference.
- Built an HLS-based training engine achieving up to 51 \times memory efficiency and 4 \times energy efficiency compared with an Nvidia RTX 3090 GPU.
- Resulting paper accepted to IEEE TCAD.

AMD-Xilinx Technology, Co-Op/Intern | Beijing, China

Jun 2023 - Sep 2023

- Developed a C++/HLS Transformer training framework with custom tensorized linear layers and nonlinear operations for LLM acceleration, achieved 30 \times ~ 52 \times saving in model size for end-to-end Transformer training.

SKILLS & RESEARCH INTERESTS

Languages & Tools Python, PyTorch, Huggingface, vLLM, C/C++, High-level Synthesis (HLS), Vivado/Vitis/XRT

ML & NLP Efficient Large Language Models (LLMs) Training/Inference, Efficient Large Reasoning Models (LRMs) (Model Compression, KV Cache Compression, Pruning, Low-rank decomposition, Early Exit, Knowledge Distillation, Quantization)

SELECTED PUBLICATIONS & PREPRINTS

Comprehensive Design Space Exploration for Tensorized Neural Network Hardware Accelerators

Jinsong Zhang, Minghe Li, **Jiayi Tian**, Jinming Lu, Zheng Zhang, under review at DAC, 2026. arXiv preprint.

SkipKV: Selective Skipping of KV Generation and Storage for Efficient Inference with Large Reasoning Models

Jiayi Tian, Seyedarmin Azizi, Yequan Zhao, Erfan Baghaei Potraghloo, Sean McPherson, Sharath Nittur Sridhar, Zhengyang Wang, Zheng Zhang, Massoud Pedram, Souvik Kundu, under review at MLSYS, 2026. arXiv preprint.

FLAT-LLM: Fine-grained Low-rank Activation Space Transformation for Large Language Model Compression

Jiayi Tian, Ryan Solgi, Jinming Lu, Yifan Yang, Hai Li, Zheng Zhang, under review at ARR Oct, 2025. arXiv preprint.

FETTA: Flexible and Efficient Hardware Accelerator for Tensorized Neural Network Training

Jinming Lu, **Jiayi Tian**, Hai Li, Ian Young, Zheng Zhang, under review at IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. arXiv preprint.

Ultra Memory-Efficient On-FPGA Training of Transformers via Tensor-Compressed Optimization

Jiayi Tian, Jinming Lu, Hai Li, Xiangwei Wang, Cong (Callie) Hao, Ian Young, Zheng Zhang, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2025.

BEBERT: Efficient and robust binary ensemble BERT

Jiayi, Tian, Chao Fang, Haonan Wang, and Zhongfeng Wang, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

Structural Pruning for Efficient LLM Inference via Low-rank Decomposition

Aug 2024 - May 2025

- Developed FLAT-LLM, a training-free, fine-grained compression method that leverages the low-rank structure of the activation space to transform and compress the model weights.
- Introduced a novel training-free rank selection algorithm that allocates ranks using a greedy redistribution strategy and can be integrated with existing low-rank LLM compression pipelines.
- Achieved strong performance on LLaMA-2, 3 and Mistral models with minimal calibration overhead (within minutes), validated across language modeling and downstream tasks.

Training Accelerator Design for Tensor-Compressed Transformer Models

Sep 2023 - May 2024

- Designed a tensor-compressed training framework for Transformer deployment on edge, significantly reducing model size, memory footprint, and computational overhead via algorithm-hardware co-design.
- Developed a fixed bidirectional contraction path and further extended it to an adaptive computing path search algorithm to improve memory and compute efficiency in long-sequence LLM training and inference.

Binary-Quantized Ensemble LLM for Fast and Robust Language Model Inference

Apr 2021 - Jun 2023

- Developed BEBERT, a novel quantization-ensemble strategy enabling efficient and accurate 1-bit BERT inference.
- Leveraged efficient knowledge distillation strategy for high training efficiency.
- Achieved $13\times$ model size reduction and $15\times$ compute savings over standard BERT with minimal accuracy loss.
- Proposed early-exit inference variant, further cutting compute by $20\% \sim 40\%$ on GLUE benchmark.