github.com/ttttttris

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.

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 VLSI Design & System Integration* | China **4.5/5.0** Fall 2019 - Fall 2023

INDUSTRIAL EXPERIENCE

Intel Corporation, *Research Intern* | Hillsboro, OR June. 2025 – Sep. 2025

- Proposed and implemented SkipKV, a training-free KV-cache compression framework featuring sentence-level selective eviction and dynamic generation control for efficient CoT reasoning.
- Designed a semantic similarity–based scoring metric to identify and remove redundant sentence spans while maintaining reasoning coherence.
- Introduced a dynamic steering mechanism to adapt hidden activations during inference, promoting concise and stable outputs.
- Demonstrated strong results on long-reasoning tasks (e.g. AIME24, LiveCodeBench) with LRMs: up to 26.7% higher accuracy vs. SoTA under equal compression, with 1.6× shorter generation and 1.7× higher throughput.

Intel Corporation, *Research Intern* | Hillsboro, OR June. 2024 - Sep. 2024

- Proposed and implemented 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× memory efficiency and 4× energy efficiency compared with an Nvidia RTX 3090 GPU.
- Resulting paper accepted to IEEE TCAD.

AMD-Xilinx Technology, *Co-Op/Intern* | Beijing, China June 2023 - Sep 2023

- Developed a C++/HLS Transformer training framework with custom tensorized linear layers and nonlinear operations for LLM acceleration, achieved 30× $\sim$ 52× 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)

PUBLICATIONS & PREPRINTS

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, 2025.

Activation-Informed Pareto-Guided Low-Rank Compression for Efficient LLM/VLM
Ryan Solgi, Parsa Madinei, **Jiayi Tian**, Rupak Swaminathan, Jing Liu, Nathan Susanj, Zheng Zhang, under review at ARR Oct, 2025. 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.

RESEARCH PROJECTS

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 models, significantly reducing model size and memory footprint.
- Developed a fixed bidirectional contraction path and an adaptive 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 - June. 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× model size reduction and 15× compute savings over standard BERT with minimal accuracy loss.
- Proposed early-exit inference variant, further cutting compute by $20\% \sim 40\%$ on GLUE benchmark.