Jiayi Tian

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

Focus on efficient training and inference of LLMs (Low-rank decomposition, Pruning, Quantization, Knowledge Distillation)

EDUCATION

University of California, Santa Barbara, *Ph.D. in Computer Engineering* | CA, USA **3.9/4.0** Fall 2023 - ongoing

Nanjing University, *B.Eng. in VLSI Design & System Integration* | China **4.5/5.0** Fall 2019 - Jun 2023

INDUSTRIAL EXPERIENCE

Intel Corporation, *Research Intern* | Hillsboro, OR June. 2025 - ongoing

- Working on efficient chain-of-thoughts reasoning via KV cache compression, leveraging methods including token eviction, KV sharing, early exit, and activation steering.

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

- Proposed a tensor-compressed LLM training accelerator using FPGA with optimized compute ordering, dataflow, and memory allocation, resulting paper accepted to IEEE TCAD.
- Achieved up to 48× memory efficiency and 3.6× energy efficiency compared to Nvidia RTX 3090.

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, TensorFlow, Huggingface, C/C++, High-level Synthesis (HLS), Vivado/Vitis/XRT
ML & NLP	Large Language Models (LLMs), Efficient Training/Inference (Model Compression, KV Cache Compression, Pruning, Low-rank decomposition, Distillation, Quantization)

PUBLICATIONS & PREPRINTS

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 July, 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, accepted to IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. arXiv preprint.

BEBERT: Efficient and robust binary ensemble BERT
Tian, Jiayi, Chao Fang, Haonan Wang, and Zhongfeng Wang, ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

RESEARCH PROJECTS

Structural Pruning for Efficient LLM Inference via Low-rank Decomposition Aug. 2024 - May. 2025

- Developed 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 - Dec. 2024

- Designed a tensor-compressed training scheme for Transformer models that reduces model size by 30 $\sim$ 52×.
- Introduced bidirectional tensor contraction to enhance memory and compute efficiency, especially in long-sequence training and inference.
- Built an HLS-based Transformer training engine achieving up to 48× memory efficiency and 3.6× energy efficiency compared with Nvidia RTX 3090.

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.