Minshen Zhang

858-214-4015 | miz055@ucsd.edu | Personal Website | LinkedIn | Github

Summary

AI infrastructure engineer focused on LLM inference optimization across GPUs and TPUs, with experience in CUDA and Pallas kernels, sparse attention, serving systems, and compiler-level profiling and benchmarking. Contributor/code owner across ByteDance TPU-Inference, AutoPallas, HiLS-Attention, and FastVideo, with first-author ML systems research.

Technical Skills

Python, C++, CUDA, PyTorch, Triton, Transformers, Megatron, Nsight, Pallas, SGLang, vLLM, XProf, Linux.

Internship Experience

Student Researcher Intern Jun. 2026 – Present
Seed Infra, ByteDance *San Jose, CA*

- **Code owner of AutoPallas**, an internal agent for automated high-performance TPU kernel optimization.
- Wrote Pallas paged attention kernel for TPU-Inference engine, reducing kernel latency from **184.81μs to 117.16μs** at 128K context length through DMA pipelining and compute scheduling; analyzed HLO/LLO and XProf traces.
- Improved QKV GEMM by **1.28×** through tile tuning; fused context normalization, quantization and output projection for **1.34× speedup**, and Add2/RMSNorm/FP8 quantization for **1.63× speedup**.
- Co-designed topology-aware parallel collectives across dtype and sharding configurations for TPU 3D-torus ICI.

Graduate Research Assistant Intern Oct. 2025 – Jun. 2026
Hao AI Lab, University of California, San Diego *La Jolla, CA*

- Code owner of FastVideo: open-source community collaboration, training infra, custom kernels, and research.
- Lead next-gen Video Sparse Attention: sparse–quantization co-design research supervised by Prof. Hao Zhang.

Machine Learning Engineer Intern, Project Lead | *Alibaba Ant Group* Jul. 2025 – Sep. 2025

- First-authored a Machine Learning Infrastructure paper (arXiv:2512.06989).

Publications & Projects

[**Project**] **AutoPallas: TPU Kernel Optimization Agent** | *ByteDance, AI Infra* Jun. 2026 – Present

- Developed a TPU kernel optimization agent backed by a comprehensive knowledge base spanning Pallas semantics, TPU compiler internals, memory layouts, DMA pipelines, and instruction scheduling.
- Built a **self-evolving knowledge base** that automatically indexes trajectories and distills reusable optimization recipes, retaining successful transformations and failure cases to guide subsequent kernel optimization.
- Achieved **1.12–3.71×** kernel speedups across multiple TPU kernel tasks compared with strong baselines, reaching up to **90.4% memory bandwidth utilization** for decode attention on TPU7x.

[**Arxiv.2607.02980**] **Hierarchical Sparse Attention Done Right** | *Tencent Hunyuan Team* Jul. 2026

- Core contributor to HiLS-Attention: native chunk-wise sparse attention that learns top-K chunk retrieval end-to-end under LM loss via landmark-token summaries and enables 64× context-length extrapolation
- Built the official SGLang serving backend for HiLS-Attention, a learned chunk-wise sparse attention scaling LLMs to 512K-token context; released as the paper's reference inference system.
- Achieved **13.5× prefill and 15.7× decode speedup** over dense attention at 512K context length.

[**Open-Source Project**] **FastVideo (4.4K Stars)** | *Hao AI Lab, AI Infra* Oct. 2025 - Jul. 2026

- Optimized LTX 2.3 inference on a single NVIDIA GB300, achieving 3× inference speed of baseline.
- Developed training infrastructure for quantization-aware distillation and world models; co-led the NVFP4 FastQAD release for consumer GPUs.

[**Arxiv:2512.06989**] **Flash Multi-Head Feed-Forward Network** | *ML Systems* Jul. 2025 – Dec. 2025

- **First-authored** a paper introducing a multi-head FFN architecture that improves LLM quality and efficiency.
- Outperformed the Llama SwiGLU baseline in perplexity and downstream evaluations, while achieving up to **1.08× speedup** and reducing peak memory usage by **3–5×**.

Education

University of California, San Diego | *M.S., CSE* | La Jolla, CA Sep. 2025 – Dec. 2026
- Teaching Assistant, CSE 291 / DSC 291: Machine Learning Systems.

University of California, Berkeley | *Exchange, EECS* | Berkeley, CA Aug. 2023 – Jan. 2024

ShanghaiTech University | *B.S., Computer Science* | Shanghai, China Sep. 2021 – Jun. 2025