

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 inference framework, serving systems, and compiler-level profiling and benchmarking. Contributor/code owner across FastAFD, Bytedance TPU-Inference, AutoPallas, HiLS-Attention, and FastVideo, with first-author ML systems research.

TECHNICAL SKILLS

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

INTERNSHIP EXPERIENCE

- Graduate Research Assistant Intern** Oct. 2025 – Present
Hao AI Lab, University of California, San Diego *La Jolla, CA*
- Contributor to **FastAFD**: LLM inference framework with Attention-MoE disaggregation for higher throughput.
 - Code owner of FastVideo: open-source collaboration, post-training infra, custom kernels, and research.
- Student Researcher Intern** Jun. 2026 – Sep. 2026
Seed Infra Compiler, ByteDance *San Jose, CA*
- Code owner of AutoPallas**, an internal agent for automated high-performance TPU kernel optimization.
 - Wrote a Pallas paged attention kernel for the TPU-Inference engine, reducing latency from **184.81 μ s to 117.16 μ s** at 128K context length through DMA pipelining and compute scheduling; analyzed HLO/LLO and XProf traces.
 - Improved QKV GEMM by **1.28 $\times$** through tile tuning; fused context normalization, quantization and output projection for **1.34 $\times$ speedup**, and Add2/RMSNorm/FP8 quantization for **1.63 $\times$ speedup**.
 - Co-designed topology-aware parallel collectives across dtype and sharding configurations for TPU 3D-torus ICI.
- Machine Learning Engineer Intern** | *Alibaba Ant Group* Jul. 2025 – Sep. 2025
- First-authored FlashFFN, a software-hardware co-design paper **accepted at NeurIPS 2026**.

PUBLICATIONS & PROJECTS

- [**NeurIPS 2026**] **Flash Multi-Head Feed-Forward Network** | *ML Systems*
- First-authored** a paper introducing a multi-head FFN architecture that improves LLM quality and efficiency.
 - Outperformed the LLM baseline in perplexity and downstream task evaluations, while achieving up to **1.08 $\times$ speedup** and reducing peak memory usage by **3–5 $\times$** .
- [**Project**] **FastAFD: High Throughput LLM Inference** | *Hao AI Lab, AI Infra*
- Improved inference throughput by up to 23.4%** in tensor-parallel settings by utilizing Reduce-Scatter and All-Gather to avoid redundant MoE token computation.
 - Implemented MiniMax-M3 **hybrid attention** support in AFD and resolved pipeline bubbles in common AFD scenarios, achieving state-of-the-art performance.
 - Integrated **speculative decoding** in the FastAFD framework, increasing throughput by 40% on ideal workloads.
 - Tuned **Expert Parallelism Load Balancing** for AF-disaggregated serving, achieving a 14% MoE-node speedup.
- [**Arxiv Preprint**] **Hierarchical Sparse Attention Done Right** | *Tencent Hunyuan Team*
- 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 $\times$ context-length extrapolation.
 - Built the official SGLang serving backend for HiLS-Attention, a learned chunk-wise sparse attention method scaling LLMs to 512K-token context; released as the paper's reference inference system.
 - Achieved **13.5 $\times$ prefill and 15.7 $\times$ decode speedup** over dense attention at 512K context length.
- [**Project**] **AutoPallas: TPU Kernel Optimization Agent** | *ByteDance, AI Infra*
- Developed a TPU kernel optimization agent with a **self-evolving knowledge base** covering Pallas, compiler internals, memory layouts, DMA pipelines, and instruction scheduling; automatically indexed optimization trajectories and distilled reusable recipes from successful transformations and failure cases.
 - Achieved **1.12–3.71 $\times$ kernel speedups** across multiple TPU kernel tasks compared with strong baselines, reaching up to **90.4% memory bandwidth utilization** for decode attention on TPU7x.

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