YICHI ZHANG

Bellevue, WA | easonzhang602@gmail.com | 434-989-0932 | github.com/Ccyest | linkedin.com/in/yichi-zhang2003

HIGHLIGHT

- Software engineer focused on **LLM inference serving and performance optimization**, with active contributions to **SGLang-Omni**.
- Hands-on with **multi-GPU inference pipelines, TTS and Omni model optimization, GPU performance benchmarking, and CI / regression infrastructure** across H100 / H200 / H20.

OPEN SOURCE & PROJECT

SGLang-Omni — Core *Contributor (Lead Higgs TTS optimization)* *April 2026 – Present*

- Led the Higgs TTS model inference-optimization workstream within SGLang-Omni: designed the optimization roadmap across the encoder, AR-decode, and vocoder stages, opened issues for the roadmap and reviewed collaborators' PRs— together delivering **[+103% throughput, +107% audio-s/s, and −51% RTF]** on **H200** (SeedTTS EN, N=1088) end-to-end at WER / speaker-similarity parity.
- Drove **CUDA Graph capture for the autoregressive decode path** (higgs-tts): tensorized per-request sampler state into batched GPU tensors, vectorized the sampler step, and integrated the graph-friendly forward into SGLang's CudaGraphRunner, removing **3–5 ms** of per-step launch overhead**.** Delivering **[+69% throughput, +74% audio-s/s, and −40% RTF]** on H200 while keep accuracy parity.
- Implemented encoder-stage optimizations — LRU caching for the audio-encoder reference path and batched audio encoding (HiggsAudioCodec.encode_batch) — cutting redundant compute and per-step overhead across the pipeline.
- Reviewed the large-scale **v0 → v1 inference-pipeline refactor**, which delivered a more maintainable architecture with **2–5× throughput** (4.7× on MMMU, 2.6× on TTS voice-cloning) and **2–4× latency reduction** at accuracy parity across MMMU, Video-MME, and SEED-TTS.
- Built CI infrastructure for **multi-card topologies** — a thinker tensor-parallel (TP=2) + disaggregated-talker stage and a staged Qwen3-Omni benchmark pipeline — with worst-of-N and hardware-aware threshold calibration and empirical memory tuning across **H20 / H100 / H200**; root-caused a silent OOM failure on H100 and contributed error-surfacing and diagnostic documentation.

SGLang — *OSS Contributor* *September 2025 – Present*

- Created and presented official SGLang tutorial videos (SGLang Diffusion, SGLang Cookbook) covering inference-pipeline walkthroughs, runtime configuration, and server deployment — helping onboard developers and accelerate community adoption.
- Expanded test coverage for SGLang's **OpenAI-compatible API endpoints** across multiple PRs, improving CPU-path reliability while keeping CPU-only unit tests resource-efficient.

WORK EXPERIENCE

Software Engineer (AMTS), Salesforce *Bellevue, WA · July 2025 – Present*

- Led and contributed to multiple Tableau Mobile end-to-end feature efforts — driving feature design and implementation, CI/CD pipelines, test-plan design, and App Store release management to ensure timely, reliable product launches.
- Partnered with PM, design, and backend teams to deliver **TabAgent**, an embedded AI assistant for Tableau that processes data for **millions+** of global users—generating summaries, answering natural-language questions, and surfacing visualization insights.
- Built a Python + LangGraph AI agent that automates the team's bug-blitz process by pre-planning test paths, analyzing DOM structure, generating and executing validation scripts, and summarizing results — improving end-to-end UX validation efficiency by **over 50%**.

Software Engineer Intern, Salesforce *Seattle, WA · May 2024 – Aug 2024*

- Implemented new Tableau-Pulse features (React Native + Redux) shipping to a Tableau Mobile release that serves business-level data management and visualization for 100k+ customers; traced internal API payloads end-to-end with Reactotron.

Software Engineer Intern, RevArt *Santa Clara, CA · Sep 2023 – Dec 2023*

- Developed an AI content assistant (ChatGPT APIs) that generates social posts from artist prompts and uploaded artwork, reducing content-creation time by **80%** and serving **10k+** artists across multiple social platforms.

AI Engineer Intern, Inspur Group Co. Ltd *Shandong, China · May 2023 – Aug 2023*

- Deployed and optimized production-grade extraction models on **cloud inference servers** backed by a scalable knowledge-graph store, and built a **LangChain + Qwen** agent to normalize heterogeneous EMR formats into structured outputs for downstream ML.

EDUCATION

University of Virginia (UVA) — B.A. Computer Science & Mathematics (Double Major) *Aug 2021 – May 2025*

- Graduated with High Distinction.

TECHNICAL SKILLS

LLM Inference & Systems: SGLang, inference serving, multi-GPU / tensor parallelism, disaggregated inference, CUDA Graph optimization, quantization, benchmarking, regression testing, RAG, LangChain / LangGraph

Infra & Tools: CUDA, Docker, Linux, Git, GitHub Actions CI/CD

Programming Languages: Python, C, C++, Go, Java, TypeScript, HTML/CSS/JavaScript, SQL, R, Arm64, x86-64