Yi Guo

(805)895-0554|guoyi0328@gmail.com|43108 Calle Sagrada, Fremont, CA, 94539|GitHub/y1guo

Skills

Languages	Python, JavaScript (Next.js, React), Rust, Java, C/C++, Swift
DB & DevOps	PostgreSQL, Redis, SQLite, Qdrant, GCP, Firebase, Docker, K8s, Git, GitHub
AI & ML	OpenAI, Anthropic, Gemini, Deepgram, Cartesia, WhisperX, XTTS, Silero-VAD, Llama3.1 Embeddings, RAG, Prompting, Few-shot Learning, Tool-use, NLP
Technologies	RESTful APIs, WebSockets, Distributed Computing, Data Visualization, GitHub Copilot

Professional Experience

RealChar, Inc. - Software Engineer Aug 2023 – Present

Project: Revia

- Reduced voice chat e2e latency to an industry-leading 500ms by adopting the architecture of self-driving cars along with a multi-agent system and frequent caching.
- Prototyped the MVP with Python, React, PostgreSQL, Firebase, Twilio and Agora.
- Major contributor of the perception, prediction, control and planner modules.
- Developed a websocket server for Agora Java SDK to communicate with the python backend.
- Designed and implemented RESTful APIs for core functionalities including calls, call histories and CRUD.

Project: Rebyte

- Identified and resolved backend latency bottlenecks by addressing issues within the Deno sandbox.
- Expanded support for additional LLM and embedding providers, integrating new tokenizers.

Project: RealChar (Open Source)

- Major contributor of key features such as RAG, phone call mode, meeting mode, and chat on image.
- Implemented RESTful API servers for WhisperX and XTTS, containerized, and deployed on GCE.
- Achieved an 8x reduction in speech-to-text latency (compared to faster-whisper) and a 4x reduction in text-to-speech latency (compared to ElevenLabs).
- Built a server-side VAD using Silero-VAD and loudness algorithms, foundational for the phone call mode.

Personal Projects

Auto-Transcribe - Open Source Developer May 2023 – Aug 2023

- Created software for automatic video transcription, incorporating stem separation and speech recognition.
- Delivered 30x real-time speed, enabling 24/7 unattended transcription with alignment.
- Developed a web UI capable of searching through 10 million audio segments by transcript within 1s.
- Ensured continuous operation with crash recovery, multi-GPU support. Transcribed 50TB of video in 1 month.
- Leveraged the tool to produce video content, garnering over 680,000 views and 1,400 engaged subscribers.

Research Experience

High-Performance Computing for Cosmology Jan 2022 – Nov 2023

- Developed high-performance code calculating cosmological parameters, achieving speeds $\sim$100x faster than existing packages like FishLSS. Improved forecasted sensitivity by $\sim$10x using advanced physics techniques.

Axion Interaction Constraints Dec 2020 – Jan 2022

- Accelerated computations by compiling Python to C (100x speedup) and by implementing distributed computing with Ray. Computed the thermal history to place leading constraints on axion-fermion interactions.

Education

University of California, San Diego La Jolla, CA

Ph.D. Physics Physics Excellence Award Sep 2018 – Dec 2023

University of California, Santa Barbara Isla Vista, CA

B.S. Physics, Mathematics Academic Honors, Worster Fellowship Sep 2014 – Jun 2018

Selected Publications

D. Green, **Y. Guo**, J. Han and B. Wallisch, "Light Fields during Inflation from BOSS and Future Galaxy Surveys," In: *JCAP* 05, p. 090 (2024) DOI: 10.1088/1475-7516/2024/05/090 arXiv: 2311.04882 [astro-ph.CO].

D. Green, **Y. Guo** and B. Wallisch, "Cosmological Implications of Axion-Matter Couplings," In: *JCAP* 02.02, p. 019 (2022) DOI: 10.1088/1475-7516/2022/02/019 arXiv: 2109.12088 [astro-ph.CO].