Yi Guo

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

Skills

Languages	Python, Next.js, React, Java, Rust, C/C++, Swift
Dev Tools	GCP, Firebase, Redis, PostgreSQL, SQLite, Qdrant, Docker, K8s, Git, GitHub, Copilot
APIs & Deploy	OpenAI, Anthropic, Gemini, Deepgram, Cartesia, WhisperX, XTTS, Silero-vad, Llama3.1
AI Skills	Embedding, RAG, Prompting, Few-shot Learning, Tool-use, NLP
Data Science	Data Visualization, Monte Carlo Simulation, Parallel/Distributed Computing

Software Development

Revia (Python, React, GCP, Redis, PostgreSQL) Jan 2024 – Present

- Reduced voice chat e2e latency to an industry-leading 500ms by adopting the architecture of self-driving cars along with a multi-agent system with heavy 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.
- Implemented restful APIs for businesses like making calls, loading replays, CRUD operations, etc.

Rebyte (Rust, React, GCP, Redis, PostgreSQL, Qdrant, RAG) Nov 2023 – Jan 2024

- Investigated the latency bottleneck of the backend and located the root cause in the Deno sandbox.
- Extended the list of supported LLM/Embedding providers, including corresponding tokenizers.

RealChar, 6K Stars (Python, React, GCP, RAG, WhisperX, XTTS, VAD) Aug 2023 – Dec 2023

- Major contributor for RAG, Phonecall mode, Meeting mode, Chat on pictures
- Implemented restful API servers for WhisperX and XTTS. Containerized and deployed on GCE.
- Reduced STT TTFT by 8X (vs faster-whisper) and TTS TTFB by 4X (vs ElevenLabs).
- Built a server side VAD as a state machine based on Silero-vad and loudness. Bedrock for Phonecall mode.

Auto-Transcribe (Python, Whisper, Demucs, Multi-process) May 2023 – Aug 2023

- Created a software to automatically transcribe videos, including stem separation and speech recognition.
- 30X real-time speed, 24/7 unattended transcription with alignment. No existing products on the market.
- Built a WebUI to search for audio segments by transcript among 10M sentences within 1 sec.
- 24/7 robustness, with crash recovery and multi-GPU support. Transcribed 50TB video within a month.
- Used it to create video content, resulting in 680K+ views and 1.4K+ engaged subscribers.

Research

Cosmological Parameters Forecast (HPC, Python, Parallel/Distributed) Jan 2022 – Nov 2023

- Developed high performance code computing $f_{\rm NL}$ $\sim$100X faster than popular packages (e.g. FishLSS).
- Increased the forecasted sensitivity by $\sim$10X with cutting-edge physics techniques.

Standard Model with Axion (HPC, Python, Parallel/Distributed) Dec 2020 – Jan 2022

- Gained the best constraints of axion-fermion interactions by computing the thermal history of early universe.
- 100X sped up by compiling native Python into C. Another 50X by distributed computing with Ray.

Spectroscopy and Photometry (ML, Python, C/C++, Fortran, Shell) Sep 2015 – Jun 2017

- Developed a spectrum fitting package (10X workflow speed up) and a galaxy photometry pipeline (100X).

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

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].