Sheng (Tim) Chen

4067 Miramar Street, San Diego CA 92037

Email: tic038@ucsd.edu Github Profile: https://timc1325.github.io/ Tel: +1 6572306683

OBJECTIVE

Driven to develop scalable machine learning tools that turn mathematical insight into practical impact.

EDUCATION

University of California, San Diego *Sep. 2024 - Present*
Ph.D. in Electrical and Computer Engineering (Machine Learning Track)

University of California, Los Angeles *Sep. 2020 - Jun. 2024*
M.S. in Applied Statistics; B.S. in Mathematics

EXPERIENCE

Uber Technologies Inc., San Francisco, California
PhD Software Engineer Intern, Platform Engineering (Capacity & Efficiency) *Jun. 2025 - Sep. 2025*

- Built a graph-based infrastructure demand forecaster, unifying historical usage and simulation into a standard record schema for apples-to-apples actuals vs. forecasts across platforms and hardware
- Shipped production SQL/Python pipelines that model growth, buffers, and efficiency, with observability and lightweight dashboards to support planning decisions and forecast stability
- Prototyped dynamic parameterization, cost modeling, and ML-guided optimization on graph-structured time-series data to inform the next iteration

University of California, San Diego (UCSD), San Diego, California
Graduate Student Researcher *Sep. 2024 - Jun. 2028*

- Optimized machine learning pipelines for efficiency, reducing computational costs through model compression and distributed training.
- Implemented privacy-preserving methods like cryptographic protocols (e.g., MPC using MP-SPDZ) for secure, privacy-preserving machine learning workflows.
- Developed data and tensor parallelism techniques for large-scale model training, incorporating zero-knowledge proofs (ZKPs) to efficiently verify correctness of distributed computations without exposing model parameters.

PROJECTS

KV Cache Compression for Efficient LLM Inference , UCSD
Project Leader *Jan. 2024 - Mar. 2024*

- Optimized ChatGLM2-6B-32k for long-context QA by implementing LLMLingua-based KV cache compression along the token dimension, enabling efficient inference on LongBench.
- Developed and analyzed token-wise compression methods (pruning and caching) to enhance the efficiency–accuracy balance of long-context language models across accuracy, latency, and memory.
- Integrated a summarization model as a preprocessing stage to mitigate out-of-memory failures and enhance scalability for extended-context inputs.

Optimization on Chip Design, UCSD
Project Member and Second Author *Sep. 2024 - Dec. 2024*

- Developed a constraint-aware multi-objective Bayesian optimization framework, achieving up to 1.7× improvement in hypervolume and eligibility rates.
- Validated the framework on HDnn-PIM designs, outperforming random search and existing methods (EHVI, NEHVI) across datasets.

PUBLICATIONS

2024 (M.S. Thesis) **Goal-Oriented Forecasting: Predicting Soccer Match Outcomes with Deep Learning**