Tianyu Xiong

Columbus, OH, USA · 614-696-3066 · tianyuxiong.zz@gmail.com · https://xtyinzz.github.io/

Summary

PhD Researcher and Engineer in applied Machine Learning (ML), specializing in building high-performance deep learning systems. Proven track record of engineering AI models with critical task-specific performance & efficiency gains and architecting AI-empowered productivity applications.

EDUCATION

Ohio State University — **Columbus, OH**
PhD, Computer Science and Engineering, Research Advisor: Dr. Han-Wei Shen — 08/2021 – 05/2026
GPA: 3.80/4.00
Honors: Distinguished University Fellowship ($56,632)

Ohio State University — **Columbus, OH**
B.S., Computer Science and Engineering — 08/2017 – 05/2021
GPA: 3.98/4.00

QUALIFICATIONS

Programming: Python, C++, Java, JavaScript, HTML, CSS, MATLAB, R, SQL
Frameworks: PyTorch, TensorFlow, Accelerate, MPI, sklearn, SciPy, Pandas, Matplotlib, Git, Docker
Concepts: Machine Learning (ML), Deep Learning (DL), Generative AI, Distributed Computing, AI Agents, Large Language Model (LLM), Vision-Language Model (VLM), Multimodal LLM

EXPERIENCE

Ohio State University — **Columbus, OH**
Machine Learning Research Associate — 08/2022 – Present

- **High-Performance AI/ML Systems**
  - Built and optimized generative models for 2D/3D data, reducing data generation time from 1 hour to 2 seconds **(a 1800x speedup)** and enabling near real-time data analysis impossible before.
  - Delivered a novel AI compression model that **cut storage costs by 99.99%** (10,000:1 compression ratio) while maintaining usable data fidelity, a tier where existing industry-standard tools fail.
  - Engineered predictive AI models with built-in **Uncertainty Quantification (UQ)** to address AI reliability, which flagged low-confidence predictions for critical review and enhanced AI reliability.
- **AI-Empowered Productivity Applications**
  - Architected an intelligent document analysis tool empowered by an **LLM backend** for technical document comprehension, with **up to +3 higher user ratings (scale of 5)** vs. alternative tools.
  - Developed a diagnostics dashboard for generative AI models, visualizing feature-space drift during training to **accelerate debugging of optimization issues** for the AI developers.

Argonne National Laboratory — **Lemont, IL**
Machine Learning Research Aide Intern — 06/2023 – 08/2023

- Conducted a comprehensive benchmark of leading AI compression models, delivering a performance analysis report that directly informed the technical roadmap for the lab's surrogate modeling project.
- Delivered on the primary internship goal of producing publishable research by rapidly prototyping novel predictive model for science accepted at IEEE VIS, the premier conference of the lab’s interest.

PROJECTS

Scalable AI Model Development System

- Engineered **a distributed ML pipeline** using PyTorch and Accelerate to scale the training of neural networks on multi-GPU nodes, reducing training time from weeks to days.
- **Solved critical memory scalability concerns during inference on massive datasets** with a chunked data processing system and a custom distributed evaluation framework with MPI collectives.
- **Validated model performance** using industry-standard metrics (PSNR, SSIM) and advanced volume rendering for rigorous quantitative and qualitative assessment.
- **Tech Stack:** Python, PyTorch, Accelerate, MPI, CUDA, Pandas, Matplotlib

LLM-Powered Document Analysis Application

- Built **a full-stack application** for LLM-empowered document understanding, which transforms raw LLM output into an interactive knowledge discovery tool for users.
- Engineered the core backend innovation: **prompt-engineering plus a core data structure** that converts LLM responses into hierarchical format, enabling users to track and synthesize insights.
- Pioneered **a multi-modal "answer engine", a design followed by the most recent Google Gemini interface**, that elevated standard text-based LLM responses by automatically sourcing and embedding relevant figures and tables from the document, presenting richer information in frontend.
- **Solved user disorientation issues** in deep-dive analyses by designing a novel tree visualization that served as a persistent knowledge map, enabling efficient navigation of complex information hierarchies**.**
- **Demonstrated system usability through user studies,** where the tool achieved significant advantage in application rating and qualitative feedback compared to other LLM-backed systems for the task**.**
- **Tech Stack:** Python, TypeScript, Flask, OpenAI API, Vue.js, D3.js, Google Firebase

PUBLICATIONS

[1] (Under Review for ICLR 2026) **Xiong, T.**, Wurster, S. W., & Shen, H. W. (2025). Refine Now, Query Fast: A Decoupled Refinement Paradigm for Implicit Neural Fields.

[2] (Under Review for CHI 2026) Qiu, R., **Xiong, T.**, Yen, P. Y., & Shen, H. W. (2025). InterDoc: Facilitate Iterative Information Seeking in Scholarly Documents via Non-linear Interaction and Adaptive Multimodal Summarization.

[3] (Under Review for ICLR 2026) Li, Z., Duan, Y., **Xiong, T.**, Chen, Y. T., Chao, W. L., & Shen, H. W. (2025). High-Fidelity Scientific Simulation Surrogates via Adaptive Implicit Neural Representations.

[4] **Xiong, T.**, Wurster, S. W., Guo, H., Peterka, T., & Shen, H. W. (2024). Regularized multi-decoder ensemble for an error-aware scene representation network. IEEE Transactions on Visualization and Computer Graphics.

[5] Wurster, S. W., **Xiong, T.**, Shen, H. W., Guo, H., & Peterka, T. (2023). Adaptively placed multi-grid scene representation networks for large-scale data visualization. IEEE Transactions on Visualization and Computer Graphics, 30(1), 965-974.

[6] Li, H., **Xiong, T.**, & Shen, H. W. (2022, October). Efficient Interpolation-based Pathline Tracing with B-spline Curves in Particle Dataset. In 2022 IEEE Visualization and Visual Analytics (VIS) (pp. 140-144). IEEE.