Yuhao Huang

✆ Phone: +1 (812) 508 1849
✉ Email: u1430219@utah.edu ❐ LinkedIn Profile

Professional Summary

Ph.D. candidate in Applied Mathematics (Machine Learning) at the Scientific Computing and Imaging (SCI) Institute, University of Utah. Earned an M.S. in Applied Mathematics from Northwestern University (IL, USA). Over four years of research and engineering experience in ML & AI, including LLM-based recommendation, diffusion and LLM models for scientific data generation, and ML applications in finance. Published at ICML, ICLR, NeurIPS, and CVPR. Skilled in machine learning, optimization, and generative modeling (e.g., diffusion, LLMs, multimodal generation) and LLM fine-tuning. Seeking a Summer 2026 internship as a Machine Learning Engineer or Researcher.

Education

❐ **University of Utah, UT, USA** Aug 2022 – Present
Ph.D. in Applied Mathematics (Machine Learning)
GPA: 3.93/4.00 | with Qualify Exam Passed

❐ **Northwestern University, IL, USA** Sep 2019 – Jan 2021
Master in Applied Mathematics (Machine Learning & Data Science Track)
GPA: 3.78/4.00 | with Thesis in *GPU-Accelerated Scientific Computing with CUDA*

❐ **Hohai University, Nanjing, China** Sep 2015 – Jun 2019
Bachelor in Information & Computing Science
GPA: 3.81/4.00 | with Thesis in *Non-negative Matrix Factorization via Heuristic Algorithm*

Technical Skills

❐ **Programming:** Python, C++/CUDA, Linux

❐ **Libraries**: PyTorch(DDP, Lightning), JAX, Tensorflow, Scikit–learn, Matlab, Pandas, NumPy, Spark, Hugging Face

❐ **Expertise**: Large Language Model Fine-tuning, Distributed Training, Generative Modeling (diffusion, LLM), Multimodal Modeling, Model Optimization.

Work Experience

❐ **VortexNet LLC**, CA May 2025 – Aug 2025
Machine Learning Intern
Enhanced product recommendation systems by fine-tuning LLMs, deploying scalable ML pipelines.

- Improved recommendation relevance by 10% in NDCG@5 through fine-tuning a large language model (LLM) with the OpenP5 framework on 200K customer click-sequence records.
- Built and deployed a production-grade end-to-end ML pipeline to pre-compute and serve personalized product recommendations, enhancing scalability and efficiency.
- Launched the LLM-based recommendation model via A/B testing, achieving an 3% increase in click-through rate (CTR) compared to the baseline system.
- Applied RLHF by learning a reward model from implicit feedback and performing KL-regularized policy optimization.

❐ **Argonne National Laboratory**, IL May 2024 – Dec 2024
Machine Learning Intern
Developed generative models (1) flow-based and (2) diffusion-LLM-based models for scientific sequential data – DNA promoter/enhancer sequence design (forecasting). Trained on datasets of 100K and 200K sequences and further improved performance by fine-tuning and speed up inference by distillation.

- Built a diffusion–LLM-based model via parameter-efficient fine-tuning of a pretrained GPN-MSATransformer with an improved fine-tuning objetive, yielding an average 5% improvement against diffusion and autoregressive baselines.
- Developed a flow-based model combining flow matching with argmax flow to handle discrete sequences, and introduced a regularized fine-tuning objective that improved accuracy by 3% on average compared to diffusion and autoregressive baselines.
- Scaled training across multiple GPUs by combining PyTorch DDP and the lab's HPC tools.

❐ **AQUMON Digital Wealth Management**, Hong Kong SAR Nov 2021 – Jul 2022
Quantitative Researcher Intern,

- Handled marketing data processing with Apache Spark to support backtesting and feature engineering for trading models.
- Developed trading models using ensemble learning techniques combining RNNs and tree-based methods.
- Designed and implemented ETF arbitrage strategies with integer programming algorithms; conducted statistical analysis to evaluate trading efficiency.
- Researched and evaluated machine learning models (RNN, ARIMA, LSTM, CNN, Attention, Transformer) for time-series forecasting applications.

Research Experience

❒ **Scientific Computing & Imaging Institute, University of Utah**, Salt Lake City, IL August 2023 – Present
Research Assistant,

✎ ***Diffusion and Flow Matching for Spatiotemporal Data (Video Prediction; Dynamics Simulation)***

Developed score-based diffusion and flow-matching models for spatiotemporal sampling in dynamical system trajectory and video prediction.

– Formulated dynamical system trajectory forecasting with score-based diffusion and flow-matching backbones.
– Developed an autoregressive latent diffusion/flow framework based on diffusion transformer and VQGAN for video prediction that conditions on recent and long-range frames, capturing momentum cues and improving denoising accuracy.
– Designed fine-tuning objectives with regularization to improve the models' performance, reducing FVD by 10% on video prediction benchmarks and achieving SOTA in dynamical system trajectory forecasting.
– Accelerated AR inference by 20× using a MeanFlow-style distillation scheme, enabling efficient one-step sampling.

✎ ***Flow-Matching for Text-to-Image Multimodal Generation via Stable Diffusion Fine-Tuning***

Built a latent flow-matching text-to-image model fine-tuned from Stable Diffusion on a 200k-image downstream dataset; Accelerated inference via distillation, and improved the distilled model with reinforcement learning.

– Built a latent flow-matching text-to-image model by fine-tuning Stable Diffusion: froze the VAE and text encoder, repurposed the UNet to predict latent velocity, and selectively fine-tuned middle and upper blocks.
– Developed one-step inference model via distillation: extended MeanFlow-style distillation to latent flow matching, enabling one-step sampling and 20× faster inference vs. standard FM and next-token transformer baselines.
– Improved a distilled generative model by combining supervised fine-tuning with feedback (SFT + RLHF), achieving a 40% reduction in one-step FID. Developed a novel likelihood-based SFT objective and implemented reinforcement learning with perceptual feature rewards.
– Training systems scaling: Deployed `PyTorch DDP` on 8× GPUs with mixed precision and gradient checkpointing, achieving 3×–4× higher throughput.

✎ ***Research on Graph Learning with Mamba for Protein Sequential Data***

This project focused on protein property classification and prediction by representing proteins as both graph-structured and sequential data, leveraging GNNs and state-space models – Mamba.

– Deigned and Developed a GNN-based local geometric encoders for protein structures
– Implemented a fragment tokenization algorithm that converts backbone geometry into sequential tensor representations.
– Integrated the resulting geometric sequences with the Mamba state-space model to capture long-range dependencies across protein chains, improving classification of protein folds and prediction of physicochemical properties.

Publications

(*: equal contribution; my name is **bolded**)

1. S-H. Wang, **Y. Huang**, T. Transue, J. Baker, J. Forstater, T. Strohmer, B. Wang. "Towards Multiscale Graph-based Protein Learning with Geometric Secondary Structural Motifs". The 39th Neural Information Processing Systems, **NeurIPS 2025**.
2. **Y. Huang**, T. Transue, S-H. Wang, W. Feldman, H. Zhang, B. Wang. "Improving Flow Matching by Aligning Flow Divergence". Proceedings of the 42nd International Conference on Machine Learning, **ICML 2025**.
3. T. Sun, **Y. Huang**, L. Shen, K. Xu, B. Wang. "Investigating the Role of Weight Decay in Enhancing Nonconvex SGD". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, **CVPR 2025**.
4. **Y. Huang***, S.-H. Wang*, J. Baker, Y.-E. Sun, Q. Tang, B. Wang. "A Theoretically-Principled Sparse, Connected, and Rigid Graph Representation of Molecules". The 13th International Conference on Learning Representations, **ICLR 2025, Oral 1.8%**.
5. **Y. Huang**, Q. Wang, A. Onwunta, B. Wang. "Efficient Score Matching with Deep Equilibrium Layers". The 12th International Conference on Learning Representations, **ICLR 2024**.