

James Song

shxjames@umich.edu shxjames.github.io github.com/shxjames

Education

University of Michigan

Expected Dec 2027

M.S. Computer Science and Engineering – Starting Fall 2026

University of Michigan

May 2026

B.S.E. Computer Science and Engineering, B.S. Mathematics

GPA: 3.94/4.00

- Coursework: Machine Learning Theory • GPU Programming • Generative Models • Foundations of LLMs • Linear Algebra • Operating Systems • Probability Theory • Discrete Stochastic Processes
- Teaching: Instructional Aide for EECS 445 (Machine Learning), Fall 2024

Publications

[Nodule-Aligned Latent Space Learning with LLM-Driven Multimodal Diffusion for Lung Nodule Progression Prediction \(NAMD\)](#)

James Song*, Yifan Wang*, Chuan Zhou, Liyue Shen

In Submission to MICCAI 2026

[When “Correct” Is Not Safe: Can We Trust Functionally Correct Patches Generated by Code Agents?](#)

Yibo Peng*, James Song*, Lei Li*, Xinyu Yang, Mihai Christodorescu, Ravi Mangal, Corina Pasareanu, Haizhong Zheng, Beidi Chen

ACL Main 2026

Research Experience

Research Assistant, UofM Department of ECE – Ann Arbor, MI

Feb 2025 – Present

Advisor: Prof. Liyue Shen

- Designed and trained NAMD, a multimodal latent diffusion model with a U-Net backbone conditioned on prior scans and MedGemma 4B clinical text embeddings via cross-attention for spatiotemporal medical image prediction, achieving 0.805 AUROC vs. 0.742 baseline, surpassing all prior methods.
- Constructed a custom VAE with a five-component loss (L1, KL, LPIPS, alignment, predictive representation) to structure the latent space by clinical attributes, leading to a 4.9% AUROC improvement in clinical diagnosis.

Research Intern, CMU InfiniAI Lab – Pittsburgh, PA

May 2025 – Oct 2025

Advisor: Prof. Beidi Chen

- Developed a black-box attack exposing Functionally Correct yet Vulnerable (FCV) patches in LLM code agents (SWE-Agent, OpenHands) evaluated in Docker-based sandboxes, achieving 18–63% attack success rates across 12 model-agent configurations on SWE-Bench while maintaining 59–94% functional correctness.

Research Assistant, UofM Department of CSE – Ann Arbor, MI

Oct 2024 – July 2025

- Investigated multi-choice reinforcement learning for LLM post-training as an alternative to DPO on Llama 3.2 1B/3B, Llama 3.1 8B, and Qwen 2.5 1.5B, using a reward model to score and weight multiple candidate responses, achieving a ~54% win rate against DPO-trained models on AlpacaEval.
- Implemented distributed LoRA DPO fine-tuning with HuggingFace Accelerate on 2× NVIDIA A40 GPUs, and multi-GPU inference for 7B–72B-parameter models (Llama-3.1-70B-Instruct, Qwen 2.5 72B) across 4× A40 GPUs.
- Built a synthetic preference dataset of ~100K question-response pairs using vLLM for high-throughput inference with Llama-3.1-70B-Instruct and Qwen 2.5 72B; benchmarked post-trained models with lm-evaluation-harness.

Research Assistant, UofM Transportation Institute – Ann Arbor, MI

May 2023 – Aug 2024

Advisor: Prof. Jingwen Hu

- Developed an automated mesh morphing pipeline using Adaptive Radial Basis Functions and Coherent Point Drift (CPD) for T1–12 vertebrae, reducing maximum geometric error by 75% and achieving sub-millimeter mean-squared error in MATLAB.
- Achieved second place in the 2023 UMTRI Student Poster Competition.

Teaching Experience

Instructional Aide (EECS 445: Machine Learning), UM CS – Ann Arbor, MI Aug 2024 – Dec 2024

- Supported student learning with a team of 12 instructional aides by providing weekly office hours and curating discussion notes and projects for 300+ students.
- Taught weekly discussion sections on topics including Kernelized Linear Regression and Classification, Deep Neural Networks, Transformers, and Generative Models.

Projects

Linear Mode Connectivity (LMC) Experiment June 2024 – July 2024

- Reimplemented GPT-2 (124M params) from scratch in PyTorch, including multi-head self-attention, layer normalization, and the full training loop with AdamW optimizer and learning rate scheduling.
- Replicated LMC experiments on CIFAR-10 across LeNet, ResNet-20, VGG-16, and decoder-only transformers; found ResNet-20 and VGG-16 exhibit loss barriers peaking at ~80% interpolation error while LeNet remains linearly connected, consistent with Frankle et al. (2020).

VidBite June 2024 – July 2024

- Built a retrieval-augmented generation (RAG) pipeline using LangChain, Claude 3.5 Sonnet, and Gemini to auto-generate math/science explainer videos; indexed Manim documentation via vector search to ground code generation in verified API usage.

Boogie Battle Sept 2024 – Sept 2024

- Built a real-time multiplayer dance battle game in Unity/C# with a Python backend performing pose estimation via OpenCV; generated photorealistic 3D avatars using NeRF and mapped skeletal keypoints to rigged character animations.

Skills

Programming Languages: Python, C++, C, MATLAB, Go

Tools and Frameworks: PyTorch, Accelerate, CUDA, NumPy, HuggingFace Diffusers, vLLM, transformers, wandb, Docker, LangChain, Scikit-Learn, Git