Jucheng Shen

js237@rice.edu | juchengshen.github.io | github.com/juchengshen

Education

Rice University Expected May 2027

B.A. in Computer Science, Math, Econ, GPA – 3.8/4.0

- Relevant Coursework: Algorithms and Data Structures, Discrete Mathematics, Computer Systems, Honors Linear Algebra, Real Analysis, Probability and Statistics, Honors ODE, Calculus on Manifolds (IP), Computer Vision (IP)

Experience

Research Intern | California Institute of Technology, advised by Prof. Anima Anandkumar Feb 2026 – Present

- Investigating neural operator learning on continuous language representations for text generation.

Research Intern | Rice University, advised by Prof. Anastasios Kyrillidis Jan 2026 – Present

- Exploring recursive reasoning models using the Looped Transformer architecture.

Research Intern | Princeton University, advised by Prof. Zhuang Liu Nov 2025 – Feb 2026

- Investigating data filtering and dataset merging strategies for large-scale text-to-image models, with a focus on improving generalization and training efficiency.
- Studying the role of image–text alignment by analyzing how VAE feature–caption consistency influences representation learning in open text-to-image foundation models.

Research Intern | UT Austin & Intel Labs, advised by Prof. Atlas Wang & Dr. Souvik Kundu May 2025 – Sep 2025

- Proposed One-Shot Dynamic Thresholding (OSDT), a training-free adaptive decoding method for diffusion LLMs achieving 24–50% higher throughput without accuracy loss.
- Led a collaboration with Intel Labs, proposing CadLLM, a unified adaptive decoding system that jointly tunes step size, block size, vocabulary, and thresholds, achieving up to 2.3× speedup on diffusion LLMs.

Research Intern | Rice University, advised by Prof. Yuke Wang Apr 2025 – Aug 2025

- Implemented *SuperGen*, a novel framework enabling generation of ultra-high resolution (2–4K) videos with limited GPU resources, on two baseline model architectures.
- Conducted end-to-end and ablation experiments, assisted with figure production and paper writing for publication.

Publications

Improving the Throughput of Diffusion-based Large Language Models via a Training-Free Confidence-Aware Calibration. Preprint 2025. **Shen, J.**, Sarkar, G., Ro, Y., Nittur Sridhar, S., Wang, Z., Akella, A., Kundu, S. arXiv:2512.07173.

Beyond Static Cutoffs: One-Shot Dynamic Thresholding for Diffusion Language Models. NeurIPS 2025 Efficient Reasoning Workshop. **Shen, J.**, Ro, Y. arXiv:2511.02077.

SuperGen: An Efficient Ultra-High-Resolution Video Generation System with Sketching and Tiling. Preprint 2025. Ye, F., Zhao, Z., Mu, Y., **Shen, J.**, Li, R., Wang, K., Sun, D., Agarwal, S., Lee, M., Cao, T., Akella, A., Krishnamurthy, A., Ng, T. S. E., Tu, Z., Wang, Y. arXiv:2508.17756.

Projects

Board Member (Tracks and Workshops) | HackRice Apr 2025 – Oct 2025

- Hosted 2 technical workshops for 100+ attendees; ran Q&A and live coding demos.
- Co-designed and shipped the official starter code used by 500+ hackers (repo).
- Designed competition tracks and judging criteria for HackRice 15.

Honors & Awards

Frank Liu Jr. Prize for Creative Innovations in Music, Fashion, and the Arts in Napier Rice Launch Challenge (link).

Technical Skills

Tools & Frameworks: Python, Numpy, PyTorch, JAX, Bash/Shell scripting, HPC/Slurm, C, Java
Skills: LLM evaluation & benchmarking, large-scale experiment management, academic writing