Xudong Shen

This version: February 25, 2026

Contact

✉ xudong.shen@u.nus.edu

🔗 Links: LinkedIn, Homepage, Google Scholar

Intro

I am passionate about scaling Reinforcement Learning for multi-modal, long-horizon, complex real-world tasks. My earlier work focused on AI fairness, robustness, safety, and governance, where I took an evaluation-driven approach: stress-testing systems and translating findings into model improvements.

Education

Ph.D. in Artificial Intelligence, National University of Singapore 2019–2024

B.A. in Naval Architecture & Ocean Engineering, Zhejiang University 2015–2019

Experience

Co-founder, **Gata**, Singapore 2024–2025

- Building decentralized inference; scaled a consumer ChatGPT data collection App to 15K+ users & 3.5M+ conv.

Research Intern, **Sea AI Lab**, Sea Limited (NYSE: SE), Singapore 2022–2024

- Developed a method to **optimize diffusion models for any differentiable objective** (e.g., diversity, aesthetics); improving over score/flow matching and RL.
- ICLR 2024 Oral; patents in US and China.

PhD, **National University of Singapore**, Singapore 2019–2024

- Developed interpretable & robust representation learning methods.
- LLM/VLM eval on capabilities, safety, scaling behavior, & in-context learning.
- Time-to-event modeling on million-scale lending panel data: modeled repayment as a survival process to forecast default risk and profitability.

Publications

Controllable Optimization for Generative Models

1. **Xudong Shen**, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, Mohan Kankanhalli, "Finetuning Text-to-Image Diffusion Models for Fairness", In ***ICLR*** (2024), **(Oral, top 1.2%)**.

TLDR: We developed a method to optimize diffusion models for any differentiable objective defined on the generated data, where score/noise prediction and RL fail. We applied it to control output diversity in text-to-image generation.

Foundation Model Evaluations & Training

2. Ian McKenzie, ..., **Xudong Shen**, ... (26 authors), "Inverse Scaling: When Bigger Isn't Better", In ***TMLR*** (2023).

TLDR: Shows when larger models consistently perform worse; analyzes failure modes.

3. Aarohi Srivastava, ..., **Xudong Shen**, ... (450 authors), "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models". In ***TMLR*** (2023).

TLDR: Large-scale eval that reveals where LLM capabilities scale well & where they don't.

4. Yizhong Wang, ..., **Xudong Shen**, ... (40 authors), "Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks". In ***EMNLP*** (2022).

TLDR: Instruction-tuning on 1.6K tasks boosts zero-shot unseen-task performance.

5. Kaustubh D Dhole, ..., **Xudong Shen**, ... (125 authors), "NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation". In ***NEJLT*** (2023).

TLDR: Stress-tested LLM robustness using 100+ natural-language augmentations.

Safety & Bias Test Suites for LLM/VLMs

6. Paul Röttger, ..., **Xudong Shen**, ... (22 authors), "MSTS: A Multimodal Safety Test Suite for Vision-Language Models", In *ArXv* (2025).

TLDR: Multimodal safety test: image+text prompts trigger more safety failures than text-only.

7. Margaret Mitchell, ..., **Xudong Shen**, ... (55 authors), "SHADES: Towards a multilingual assessment of stereotypes in large language models", In ***NAACL*** (2025).

TLDR: Probes multilingual stereotypes and its cross-lingual transfer in LLMs.

Robust & Interpretable Representations

8. **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, "Fair Representation: Guaranteeing Approximate Multiple Group Fairness for Unknown Tasks". In ***IEEE Trans. PAMI*** (2023).

TLDR: Learns representation with robustness guarantees that transfer to unseen tasks.

9. Ziwei Xu, **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, "Unsupervised Motion Representation Learning with Capsule Autoencoders". In ***NeurIPS*** (2021).

TLDR: Learns representations with built-in interpretability via capsule networks.

Predictive Modeling for Real-World Decision Making

10. **Xudong Shen**, Tianhui Tan, Tuan Q. Phan, Jussi Keppo, "Gender Animus Can Still Exist Under Favorable Disparate Impact: a Cautionary Tale from Online P2P Lending". In ***FAccT*** (2023).

TLDR: Time-to-event modeling to predict default & profitability on million-scale lending data.

Regulatable AI: Policy & Technical Mechanisms

11. **Xudong Shen**, Hannah Brown, Jiashu Tao, Martin Strobel, Yao Tong, Akshay Narayan, Harold Soh, Finale Doshi-Velez, "Directions of Technical Innovation for Regulatable AI Systems", In ***Communications of the ACM*** (2024).

TLDR: maps technical mechanisms that make AI easier to regulate in practice.

12. Ayse Gizem Yasar, Andrew Chong†, Evan Dong†, Thomas Krendl Gilbert†, Sarah Hladikova†, Roland Maio†, Carlos Mougan†, **Xudong Shen**†, Shubham Singh†, Ana-Andreea Stoica†, Savannah Thais†, "Integration of Generative AI in the Digital Markets Act: Contestability and Fairness from a Cross-Disciplinary Perspective", (2024), **LSE working papers series**.

TLDR: analyzes how GenAI interacts with platform regulation.

Fellowships

DAAD AInet Fellow, *DAAD*, Germany	2024
Master Kong Dream Scholarship Program, *Waseda University*, Japan	Sep. 2018-Feb. 2019
Globalink Research Internship, *York University*, Canada	May.-Aug. 2018
Erasmus+ Student Mobility, *Università di Trento*, Italy	Feb.-Jun. 2017