Xuwang Yin

● xuwangyin@gmail.com ● San Francisco, CA, US
● LinkedIn ● Github ● Google Scholar ●(434) 327-0325

Education

08/2016 - 12/2022

Ph.D, Computer Engineering, University of Virginia, United States

Ph.D thesis: Generative Modeling With Adversarial Training
Supervisor: Prof. Gustavo Rohde
Research focus: adversarial machine learning, energy-based generative models, multimodal models, optimal transport in signal processing

2011 - 2014

M.Eng, Computer Engineering, University of Science and Technology Beijing, China

Master thesis: Robust Text Detection in Natural Scene Images
Supervisor: Prof. Xu-Cheng Yin

2007 - 2011

B.Eng, Computer Science and Engineering, University of Science and Technology Beijing, China

Employment

04/2025 - Present

Independent Researcher (San Francisco/Qingdao)

- Developed a novel adversarial training framework for energy-based models enabling joint image generation and classification in a single model
- First to scale energy-based models to ImageNet 256×256 with state-of-the-art sample quality and 5-29× faster inference than diffusion models
- Single model supports generation, classification, OOD detection, and counterfactual explanations
- Paper accepted to ICLR 2026 (arXiv:2510.13872)

01/2023 - 04/2025

Research Engineer, Center for AI Safety, San Francisco, United States

- Trained harmful textual content detection model with LoRA + SFT + linear probe
- Developed model evaluation library with transformers, litellm and vllm [Utility Engineering]
- Evaluating base and chat LLMs using HELM, MACHIAVELLI, lm-evaluation-harness on capability benchmarks and safety benchmarks [Safetywashing]
- Adversarial training of LLMs against GCG jailbreak prompts [HarmBench]
- Implemented adversarial attacks against open weights multimodal LLMs [HarmBench]
- Model steering using control vector computed from embedding features of LLMs [Representation Engineering]
- Implemented latent variable-based adversarial attacks, integrated state-of-the-art vision models (CLIP, DINO, Vision Transformer, ConvNeXt) into codebase and conducted extensive model evaluation [Unforeseen Adversaries]

2014 - 2015

Computer Vision Engineer, Yuanli Education Technology, Beijing, China

Developed neural network models for detecting and recognizing text in images

Research Interests

Energy-based generative models

- Scalable training of energy-based models via adversarial training
- Text-to-image generation with energy-based models
- Hybrid discriminative-generative models

Machine learning robustness and security

- Training adversarially robust image models and LLMs
- Safety evaluation/red-teaming/jailbreaking LLMs
- Detecting adversarial examples

Selected Publications

- Joint Discriminative-Generative Modeling via Dual Adversarial Training
  Xuwang Yin, Claire Zhang, Julie Steele, Nir Shavit, Tony T. Wang
  ICLR 2026 [arxiv]

- Learning Globally Optimized Language Structure via Adversarial Training
  Xuwang Yin
  Working paper, September, 2023 [arxiv]

- Learning Energy-Based Models With Adversarial Training
  Xuwang Yin, Shiying Li, Gustavo K. Rohde,
  European Conference on Computer Vision (**ECCV 2022**) [code] [arxiv] [poster] [video]

- GAT: Generative Adversarial Training for Adversarial Example Detection and Robust Classification
  Xuwang Yin, Soheil Kolouri, Gustavo K. Rohde
  International Conference on Learning Representations (**ICLR 2020**) [code][arxiv]

- OBJ2TEXT: Generating Visually Descriptive Language from Object Layouts. [code][Link]
  Xuwang Yin and Vicente Ordonez.
  Empirical Methods in Natural Language Processing (**EMNLP 2017**, oral presentation)

- Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs [code][arxiv]
  Mantas Mazeika, **Xuwang Yin**, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, Dan Hendrycks
  NeurIPS 2025

- RenderAttack: Hundreds of Adversarial Attacks Through Differentiable Texture Generation
  Dron Hazra, Alex Bie, Mantas Mazeika, **Xuwang Yin**, Andy Zou, Dan Hendrycks, Maximilian Kaufmann
  NeurIPS 2024 Workshop on New Frontiers in Adversarial Machine Learning

- HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal [arxiv] [website]
  Mantas Mazeika, Long Phan, **Xuwang Yin**, Andy Zou, Zifan Wang, Norman Mu, Ellie Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, Dan Hendrycks

International Conference on Machine Learning (**ICML 2024)**

- End-to-End Signal Classification in Signed Cumulative Distribution Transform Space [arxiv]
  AMH Rubaiyat, Shiying Li, **Xuwang Yin**, Mohammad Shifat E Rabbi, Yan Zhuang, Gustavo K. Rohde
  IEEE Transactions on Pattern Analysis and Machine Intelligence (**PAMI 2023**)

- A Digital Camera-based Eye Movement Assessment Method for NeuroEye Examination [link]
  Mohamed Abul Hassan, **Xuwang Yin***, Yan Zhuang*, Chad M Aldridge, Timothy McMurry,
  Andrew M Southerland, Gustavo K Rohde
  IEEE Journal of Biomedical and Health Informatics, 2023

- Radon Cumulative Distribution Transform Subspace Modeling for Image Classification
  M Shifat-E-Rabbi, **Xuwang Yin**, Abu Hasnat Mohammad Rubaiyat, Shiying Li, Soheil Kolouri, Akram Aldroubi, Jonathan M. Nichols, Gustavo K. Rohde
  Journal of Mathematical Imaging and Vision, 2021 [code][Link]

- Neural Networks, Hypersurfaces, and Radon Transforms
  Soheil Kolouri, **Xuwang Yin**, Gustavo K. Rohde
  IEEE Signal Processing Magazine, 2020 [code][IEEEXplore link][arxiv]

- Robust Text Detection in Natural Scene Images.
  Xu-Cheng Yin, **Xuwang Yin**, Kaizhu Huang, Hong-Wei Hao.
  IEEE Transactions on Pattern Analysis and Machine Intelligence (**PAMI 2014**) [link]

- Effective text localization in natural scene images with MSER, geometry-based grouping and AdaBoost.
  Xuwang Yin, Xu-Cheng Yin, Hong-Wei Hao, Khalid Iqbal.
  International Conference on Pattern Recognition (**ICPR 2012**) [IEEEXplore link]

- Representation Engineering: A Top-Down Approach to AI Transparency [arxiv]
  Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, **Xuwang Yin**, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, Dan Hendrycks

- Testing Robustness Against Unforeseen Adversaries [arxiv] [github]
  Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, **Xuwang Yin**, Mantas Mazeika, Akul Arora, Adam Dziedzic, Franziska Boenisch, Tom Brown, Jacob Steinhardt, Dan Hendrycks

Awards

- Top Reviewer of NeurIPS 2022
- International Conference on Document Analysis and Recognition (ICDAR2013) robust reading competition winner (out of 15 research teams from 7 countries)
- Outstanding Graduate of Beijing, Beijing Municipal Commission of Education, Jan 2014.

Professional Service

ACL-IJCNLP 2021, EMNLP 2021, ACL 2022, ICLR 2022, ECCV 2022, NeurIPS 2022, NeurIPS 2024, ACL 2024, NeurIPS 2025, ICLR 2026, Neurocomputing journal