MEHRDAD SABERI

merhdads@gmail.com msaberi@umd.edu
Google Scholar mehrdadsaberi.github.io mehrdads

EDUCATION

Ph.D. in Computer Science Jan 2023 – Present
University of Maryland, College Park
Advisor: Soheil Feizi

M.Sc. in Computer Science Jan 2023 – Dec 2024
University of Maryland, College Park

B.Sc. in Computer Engineering Oct 2018 – Jun 2022
Sharif University of Technology
Rank 3rd

RESEARCH INTERESTS

- **Content authenticity:** AI-generated media detection, watermarking, and data provenance.
- **Trustworthy generative AI:** robustness of text-to-image, video, and diffusion models.
- **Interpretability & evaluation:** failure analysis of generative and vision-language models.
- **Model unlearning:** data influence and efficient unlearning in large-scale models.

RESEARCH EXPERIENCE

Research Intern, **CAI Team, Adobe** May 2025 – Dec 2025
Worked on robust content authenticity and image watermarking for AI-generated media.
Mentor: John Collomosse

Research Intern, **GenAI Team, Cruise** Jun 2024 – Nov 2024
Developed video inpainting models for synthetic data generation to train autonomous vehicle perception systems on rare events.
Mentor: Ashish Shrivastava

Research Assistant, **Reliable AI Lab, UMD** Jan 2023 – Present
Research on AI-generated media and text detection, robust image watermarking and data provenance, interpretability and failure analysis of text-to-image and vision-language models, and model unlearning in large language models.
Mentor: Soheil Feizi

Summer Research Intern, **Theory of Machine Learning Lab, EPFL** Jul 2021 – Sep 2021
Studied adversarial robustness under non-L_p perturbations (e.g., Wasserstein, LPIPS) and used semantic adversaries to improve adversarial training.
Mentor: Nicolas Flammarion

Research Assistant, **Robust and Interpretable Machine Learning Lab, SUT** Oct 2020 – May 2021
Worked on preventing catastrophic overfitting in models trained with fast single-step adversarial training.
Mentor: Mohammad Hossein Rohban

Research Intern, **Max-Planck-Institut für Informatik** Jun 2020 – Sep 2020
Designed black-box transformations that reinforce routing schemes against independently distributed node failures while preserving original routing policies.
Mentor: Christoph Lenzen

HONORS AND AWARDS

- Silver Medal, **International Olympiad in Informatics (IOI)** 2018.
- Gold Medal (Rank 1), **Iranian National Olympiad in Informatics** 2017.
- Silver Medal, **Iranian National Olympiad in Informatics** 2016.
- Silver Medal, **AITMO Regional Competition** 2013.
- Gold Medal, **ICPC** Regional Contest 2019.
- **Grandmaster** on Codeforces (profile), 2018.

PUBLICATIONS

Watermarking and Content Authenticity

- **Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks**
  Mehrdad Saberi, *Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, Soheil Feizi*
  💡 Featured in Wired, The Register, ArsTechnicia articles, PrivacyCon 2024, and AGI Leap Summit 2024.
  ICLR 2024

  ,
- **DREW: Towards Robust Data Provenance by Leveraging Error-Controlled Watermarking**
  Mehrdad Saberi, *Vinu Sankar Sadasivan, Arman Zarei, Hessam Mahdavifar, Soheil Feizi*
  Arxiv Preprint

- **IConMark: Robust Interpretable Concept-Based Watermark for AI Images**
  Vinu Sankar Sadasivan, ***Mehrdad Saberi****, Soheil Feizi*
  GenAI Watermarking Workshop @ ICLR 2025

- **Adversarial Paraphrasing: A Universal Attack for Humanizing AI-Generated Text**
  Yize Cheng, Vinu Sankar Sadasivan, ***Mehrdad Saberi****, Shoumik Saha, Soheil Feizi*
  Neurips 2025

- **Erasing the Invisible: A Stress-Test Challenge for Image Watermarks**
  Mucong Ding, Tahseen Rabbani, Bang An, Souradip Chakraborty, Chenghao Deng, ***Mehrdad Saberi****, et al.*
  NeurIPS 2024 Competition Track

- **A Technical Report on "Erasing the Invisible": The 2024 NeurIPS Competition on Stress Testing Image Watermarks**
  Mucong Ding, Bang An, Tahseen Rabbani, Chenghao Deng, Anirudh Satheesh, Souradip Chakraborty, ***Mehrdad Saberi****, et al.*
  NeurIPS 2025 Datasets and Benchmarks

- **Securing the Future of GenAI: Policy and Technology**
  Mihai Christodorescu, Ryan Craven, Soheil Feizi, Neil Gong, Mia Hoffmann, Somesh Jha, Zhengyuan Jiang, ***Mehrdad Saberi****, et al.*
  Cryptology ePrint Archive

Interpretability and Evaluation of Generative Models

- **PRIME: Prioritizing Interpretability in Failure Mode Extraction**
  Keivan Rezaei[*]*,* ***Mehrdad Saberi***[*]*, Mazda Moayeri, Soheil Feizi*
  ICLR 2024

- **Model State Arithmetic for Machine Unlearning**
  Keivan Rezaei, **Mehrdad Saberi***, *Abhilasha Ravichander, Soheil Feizi*
  Arxiv Preprint

- **Benchmarking Text-Guided Image Editing Methods**
  Samyadeep Basu, ***Mehrdad Saberi****, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, Soheil Feizi*
  Arxiv Preprint

- **Mitigating Compositional Failures in Text-to-Image Models with Causal Text Embedding Refinement**
  Arman Zarei, Keivan Rezaei, Samyadeep Basu, ***Mehrdad Saberi****, Mazda Moayeri, Priyatham Kattakinda, Adrienne Raglin, Anjon Basak, Soheil Feizi*
  PerCom Workshops 2025

Adversarial Robustness and Algorithms

- **ZeroGrad: Costless Conscious Remedies for Catastrophic Overfitting in the FGSM Adversarial Training**
  Zeinab Golgooni, ***Mehrdad Saberi****, Masih Eskandar, Mohammad Hossein Rohban*
  Intelligent Systems with Applications, 2023

- **Robust Routing Made Easy: Reinforcing Networks Against Non-Benign Faults**
  Christoph Lenzen, Moti Medina, ***Mehrdad Saberi****, Stefan Schmid*
  IEEE/ACM Transactions on Networking, 2023

INDUSTRY AND OTHER EXPERIENCE

Data Scientist, **Charkh** Sep 2022 – Dec 2022
Designed and deployed ML-based recommender systems for large-scale online shopping.

Algorithm Course Writer, **Quera** Oct 2018 – Dec 2018
Authored algorithm and data-structure problem sets and educational content for competitive programming.

TEACHING AND MENTORING

- **Teaching Assistant**, University of Maryland
  CMSC 250: Discrete Structures (Spring 2023).
- **Teaching Assistant**, Sharif University of Technology
  Machine Learning, Artificial Intelligence, Linear Algebra, Design of Algorithms, Discrete Structures, Advanced Programming (2019–2021).
- **Instructor**, Iranian National Olympiad in Informatics Summer Camp (2019).
- **Instructor**, Algorithms, Combinatorics and Graph Theory, Shahid Beheshti High School (2017–2018).

SKILLS

ML Libraries	PyTorch, TensorFlow, PyTorch Lightning, DeepSpeed, Hugging Face (Transformers, Diffusers)
Algorithms & Competitive Programming	Statistics, Graph theory, Combinatorics, Linear Algebra, Data Structures
Programming	Python, C++, Java, Go, Bash
Tools	Git, Docker, Linux, SQL, MongoDB, Weights & Biases , Milvus