

YUCHEN YANG

+1-410-350-6041 yc.yang@jhu.edu www.cs.jhu.edu/~yuchen413

EDUCATION

- Ph.D. Johns Hopkins University** 2021 – 2025
Computer Science
Advisor: Dr. Yinzhi Cao MD, United States
- M.S. Johns Hopkins University** 2019 – 2021
Security Informatics
Advisor: Dr. Yinzhi Cao MD, United States
- B.E. Shandong University** 2015 – 2019
Software Engineering Shandong, China

INTERESTS

Trustworthy AI, LLM/VLM Safety, Social Cybersecurity, Privacy-preserving Machine Learning

PUBLICATIONS

PREPRINTS

- 2024 **CertPHash: Towards Certified Perceptual Hashing via Robust Training**
Yuchen Yang, Qichang Liu, Christopher Brix, Huan Zhang, Yinzhi Cao
Under review, 2024
- Jailbreaking Safeguarded Text-to-Image Models via Large Language Models**
Zhengyuan Jiang, Yuepeng Hu, **Yuchen Yang**, Yinzhi Cao, Neil Gong
Under review, 2024
- Pseudo-Probability Unlearning: Towards Efficient and Privacy-Preserving Machine Unlearning**
Zihao Zhao, Yijiang Li, **Yuchen Yang**, Wenqing Zhang, Nuno Vasconcelos, Yinzhi Cao
Under review, 2024

PEER-REVIEWED CONFERENCE PAPERS (* INDICATES CO-FIRST AUTHORS)

- 2024 **SneakyPrompt: Jailbreaking Text-to-image Generative Models**
Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, Yinzhi Cao
In the Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2024
Media Coverage (Selected): MIT Technology Review [↗](#)
- Follow the Rules: Reasoning for Video Anomaly Detection with Large Language Models**
Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, Shao-Yuan Lo
In the Proceedings of the European Conference on Computer Vision (ECCV), 2024
- SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models**
Xinfeng Li*, **Yuchen Yang***, Jiangyi Deng*, Chen Yan, Yanjiao Chen, Xiaoyu Ji, Wenyan Xu
In the Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2024
- RippleCOT: Amplifying Ripple Effect of Knowledge Editing in Language Models via Chain-of-Thought In-Context Learning**
Zihao Zhao, **Yuchen Yang**, Yijiang Li, Yinzhi Cao
In the Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing (EMNLP), 2024
The first author finished the paper mainly under my mentoring

- 2023 **PrivateFL: Accurate, Differentially Private Federated Learning via Personalized Data Transformation**
Yuchen Yang*, Bo Hui*, Haolin Yuan*, Neil Gong, Yinzhi Cao
In the proceedings of USENIX Security Symposium (Usenix), 2023
Earned all Artifact Badges: Artifacts Available, Artifacts Functional, Results Reproduced
- Fortifying Federated Learning against Membership Inference Attacks via Client-level Input Perturbation**
Yuchen Yang, Haolin Yuan, Bo Hui, Neil Gong, Neil Fendley, Philippe Burlina, Yinzhi Cao.
In the proceedings of IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2023
- 2022 **Addressing Heterogeneity in Federated Learning via Distributional Transformation**
 Haolin Yuan*, Bo Hui*, **Yuchen Yang***, Philippe Burlina, Neil Gong, Yinzhi Cao
In the proceedings of the European Conference on Computer Vision (ECCV), 2022
- 2021 **Practical Blind Membership Inference Attack via Differential Comparisons**
 Bo Hui*, **Yuchen Yang***, Haolin Yuan*, Philippe Burlina, Neil Gong, Yinzhi Cao
In the proceedings of Network & Distributed System Security Symposium (NDSS), 2021





EXPERIENCES

- Johns Hopkins University**, Research Assistant MD, United States
2020.03 - current
- Working on trustworthy AI/ML advised by [Dr. Yinzhi Cao](#)
 - Research outcomes: publications at top-tier security conferences—S&P, Usenix, CCS, and NDSS, and vision/NLP conferences such as ECCV and EMNLP
- Honda Research Institute**, Research Internship CA, United States
2023.10 - 2024.02
- Working on LLMs for video anomaly detection advised by [Dr. Shao-yuan Lo](#), [Dr. Kwonjoon Lee](#), and [Dr. Behzad Dariush](#)
 - Research outcomes: a publication at ECCV 2024 and filing of a U.S. patent
- Shandong Univeristy**, Research Assistant Shandong, China
2018.12 - 2019.06
- Working on an NLP-driven auto-grading system for Chinese composition in primary and secondary education advised by [Dr. Yuqing Sun](#)
 - Research outcomes: my undergraduate dissertation: Automatic Grading System for Chinese Composition via Bi-directional LSTM
- Chinese Academy of Sciences**, Research Internship Beijing, China
2018.06 - 2018.09
- Working on developing SVM+ by distinguishing the privileged vectors within SVM (support vector machine) algorithm advised by [Dr. Yingjie Tian](#)
 - Research outcomes: a patent filed under IP Australia

AWARDS AND HONORS

- Selected** as one of three institutional nominees for Apple Scholars in AI/ML PhD Fellowship 2022.09
- Selected** to attend Individualized Cybersecurity Research Mentoring Workshop (iMentor) 2021.11
- Academic Scholarship**, Shandong University 2018.09

MEDIA COVERAGE

- Text-to-image AI models can be tricked into generating disturbing images, [MIT Technology Review](#)  2023.11
- AI art generators can be fooled into making NSFW images, [IEEE Spectrum](#)  2023.11
- Nonsense prompts trick AIs into producing NSFW images, [Technology Networks](#)  2023.11
- Researchers reveal vulnerabilities in AI models, prompting concerns, [Cryptopolitan](#)  2023.11

AI generators can be tricked into making NSFW content, JHU Engineering School News	2023.11
SneakyPrompt: Revealing the vulnerabilities of text-to-image AI, The Johns Hopkins News-Letter	2023.12

PROFESSIONAL SERVICES

CONFERENCE/JOURNAL REVIEWING

PC , ACM Conference on Computer and Communications Security (CCS)	2025
ACM Workshop on Adaptive and Autonomous Cyber Defense (AACD)	2024
AEC , IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)	2024
Reviewer , International Conference on Learning Representations (ICLR)	2025
IEEE Transactions on Dependable and Secure Computing (TDSC)	2023, 2024
IEEE Transactions on Information Forensics & Security (T-IFS)	2024
External Reviewer , IEEE Symposium on Security and Privacy (S&P)	2025
USENIX Security Symposium	2023, 2024
ACM Conference on Computer and Communications Security (CCS)	2022
ACM ASIA Conference on Computer and Communications Security (ASIACCS)	2024
IEEE Computer Security Foundations Symposium (CSF)	2022, 2024
IEEE International Conference on Distributed Computing Systems (ICDCS)	2022

ORGANIZING AND CHAIRING

Session Chair , IEEE Workshop on Deep Learning Security and Privacy (DLSP)	2024
---	------

TEACHING AND ADVISING

TEACHING

Teaching assistant, Web Security	Fall 2020, 2022
----------------------------------	-----------------

ADVISING

Zihao Zhao, undergraduate student at JHU	2023.12 - 2024.09
Qichang Liu, undergraduate student at Tsinghua University	2024.06 - 2024.09
Yichen Li, undergraduate student at Beijing University	2024.06 - 2024.09
Adila Abudurehman, master student at JHU. Now: Software engineer at Microsoft	2023.10 - 2024.02
Manoj Valeti, master student at JHU. Now: Software engineer at AWS Amazon	2023.10 - 2024.02
Amodini Vardhan, master student at JHU. Now: Security engineer at Twilio	2023.10 - 2024.02
Anning Li, undergraduate student at UESTC. Now: Master student at CMU	2023.06 - 2023.09
Ashi Garg, master student at JHU. Now: Research associate at HLTSCO	2022.09 - 2023.03

PATENT

2024	System and Method Using Reasoning for Video Anomaly Detection with Large Language Models Yuchen Yang , Kwonjoon Lee, Shao-Yuan Lo, Behzad Dariush <i>Under US Patent Application (Filed), 2024</i>
2018	A New System for Stock Volatility Prediction by Using Privileged Support Vector Machines Yuchen Yang , Zheng Yan, Haoyang Li, Zhixuan Lv, Weiwei Zhao, Simiao Zhao <i>Under IP Australia Application (No.2018101304), 2018</i>

INVITED TALKS

"Diagnose, Correct, Steer: Towards Functional, Trustworthy AI", Monash University, Invited talk	2024.10
"Zero-shot Video Anomaly Detection: Steering LLMs for Rules-based Reasoning", Voxel51, Invited talk	2024.11