YUN-YUN (ALICE) TSAI

Ph.D. Candidate in Computer Science | Columbia University

yt2781@columbia.edu | yunyuntsai.github.io | +1 (917) 498-2709

RESEARCH INTERESTS

My research focuses on trustworthy and secure AI systems, with an emphasis on developing robust and reliable machine learning models under real-world deployment challenges. My research interests span adversarial robustness and test-time adaptation, model provenance and fingerprinting, generative AI safety and evaluation, and the security and reliability of agentic AI systems.

EDUCATION

Ph.D. in Computer Science *September 2021 – Present*
Columbia University, New York, NY *Expected Graduation: 2027*
- Advisor: Prof. Junfeng Yang
- Dissertation: Test-time Defenses with Reasoning for Safe and Robust Modern AI Systems

M.Phil. in Computer Science *2025*
Columbia University, New York, NY *Degree Conferred in 2025*

M.S. in Computer Science *September 2018 – June 2020*
National Tsing Hua University, Hsinchu, Taiwan
- Advisors: Prof. Tsung-Yi Ho and Dr. Pin-Yu Chen
- Master's Thesis: Adversarial Reprogramming for Black-Box Machine Learning Models

B.S. in Computer Science *September 2014 – June 2018*
National Tsing Hua University, Hsinchu, Taiwan

WORK EXPERIENCE

Research Scientist Intern *Summer 2025*
Meta Superintelligence Labs, Bellevue, WA
- Mentors: Dr. Sinong Wang, Dr. Yuning Mao, and Dr. Shiqi Wang
- Project: Execution-Based Benchmarking and Evaluation of LLM-Generated Web Applications

- Developed WebUIProof, an execution-based benchmark that uses UI agents to evaluate the functional correctness of LLM-generated web applications. - Proposed VisRL, a reinforcement learning method that leverages a UI-agent harness and composite rewards from visual, runtime, and interaction feedback to improve WebUI generation.

Research Scientist Intern *Summer 2024*
Meta GenAI Team, New York, NY
- Mentors: Dr. Laurens van der Maaten and Dr. Chuan Guo
- Project: Robust Black-Box Fingerprinting for LLM Provenance Verification

- Developed a gradient-guided fingerprinting method that identifies rare input–output behaviors of LLMs, enabling black-box model identification and ownership verification robust to fine-tuning, LoRA adaptation, quantization, and system prompt changes.

Applied Scientist Intern *Summer 2023*
Amazon Lab126, Bellevue, WA
- Mentors: Dr. Fu-Chen Chen and Prof. Min Sun
- Project: Generalized Out-of-Distribution Adaptation with Diffusion Models (GDA)

- Developed a diffusion-based test-time adaptation method that leverages structured guidance to transform out-of-distribution inputs toward the training distribution, improving visual recognition robustness under distribution shifts.

Visiting Research Scholar *March 2019 – August 2019*
Department of ECE, University of Florida, Gainesville, FL
- Advisor: Prof. Yier Jin

- Project: Black-Box Model Extraction Attacks on Deep Neural Networks. Investigated model extraction attacks that leverage adversarially crafted queries to reconstruct the decision boundaries of victim DNNs.

Research Intern *July 2017 – June 2018*
Microsoft Cloud & AI Team, Taipei, Taiwan
• Mentor: Cheryl Hsu
• Project: Developed Cloud-Based IoT Solutions. IoT prototypes leveraging Microsoft Azure IoT Cloud and Edge computing technologies across smart city, retail, smart home, and manufacturing applications.

PUBLICATIONS

Preprints and Pending Papers

[P1] **Yun-Yun Tsai**, Yuning Mao, Shiqi Wang, Junfeng Yang, and Sinong Wang, "WebUIProof: Benchmarking WebUI Code Generators with UI-Agent Execution Harness," *arXiv preprint*, 2026.

[P2] Chunyi Wang, Yunfei Ke, Junfeng Yang, **Yun-Yun Tsai**, Penghui Li, "ARQ: Agentic CodeQL Query Refinement for C/C++ Vulnerability Detection," *arXiv preprint*, 2026.

[P3] **Yun-Yun Tsai**, Jia-Hao Liang, Chuan Guo, Junfeng Yang, Laurens van der Maaten, "RAFP: Identifying LLM Lineages via Rare-Region Fingerprints," *arXiv preprint*, 2025.

Conference and Workshop Publications

[C1] Jincen Song, Luke Tenyi Wang, **Yun-Yun Tsai**, Penghui Li, and Junfeng Yang, "Reasoning under Vision: Understanding Visual-Spatial Cognition in Vision-Language Models for CAPTCHA," in Findings of the Association for Computational Linguistics: (**AACL**) 2026.

[C2] **Yun-Yun Tsai**, Qingyuan Liu, Ruijian Zha, Victoria Li, Pengyuan Shi, Chengzhi Mao, Junfeng Yang, "REVEAL: Robust Evolution of Vision-Language Models for Explainable AI-Video Detection," (**ICCV**) 2025, Knowledge-Intensive Multimodal Reasoning Workshop

[C3] **Yun-Yun Tsai**, Fu-Chen Chen, Albert Y.C. Chen, Junfeng Yang, Che-Chun Su, Min Sun, Cheng-Hao Guo, "GDA: Generalized Diffusion for Robust Test-time Adaptation," in Proceeding of the IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR**) 2024.

[C4] **Yun-Yun Tsai**, Ruize Xu, Chengzhi Mao, Junfeng Yang, "From Detection to Deception: Are AI-Generated Image Detectors Adversarially Robust?," **CVPR** 2024, Responsible Generative AI Workshop.

[C5] Qingyuan Liu, Pengyuan Shi, **Yun-Yun Tsai**, Chengzhi Mao, Junfeng Yang, "Turns Out I'm Not Real: Towards Robust Detection of AI-Generated Videos," **CVPR** 2024, Generative Models for Computer Vision workshop.

[C6] **Yun-Yun Tsai**, Chengzhi Mao, Junfeng Yang, "Convolutional Visual Prompt for Robust Visual Perception," 37th Conference on Neural Information Processing Systems (**NeurIPS**) 2023.

[C7] Lei Hsiung, **Yun-Yun Tsai**, Pin-Yu Chen, and Tsung-Yi Ho, "Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations," in Proceeding of the IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR**) 2023.

[C8] **Yun-Yun Tsai**, Ju-Chin Chao, Albert Wen, Zhaoyuan Yang, Chengzhi Mao, Tapan Shah, Junfeng Yang, "Test-time Detection and Repair of Adversarial Samples via Masked Autoencoder," in Proceeding of the IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR**) 2023, AdvML workshop

[C9] Lei Hsiung, **Yun-Yun Tsai**, Pin-Yu Chen, and Tsung-Yi Ho, "CARBEN: Composite Adversarial Robustness Benchmark," in Proceeding of International Joint Conference on Artificial Intelligence (**IJCAI**), 2022.

[C10] Chao-Han Huck Yang, **Yun-Yun Tsai**, and Pin-Yu Chen "Voice2Series: Reprogramming Acoustic Models for Time Series Classification," in Proceeding of International Conference on Machine Learning (**ICML**), 2021.

[C11] **Yun-Yun Tsai**, Pin-Yu Chen, and Tsung-Yi Ho, "Transfer Learning without Knowing, Reprogramming black box machine learning model with scarce data and limited resources," in Proceeding of International Conference on Machine Learning (**ICML**), 2020.

[C12] Honggang Yu, Kaichen Yang, Teng Zhang, **Yun-Yun Tsai**, Tsung-Yi Ho, Yier Jin, "CloudLeak: Large-Scale Models Extraction Through Adversarial Examples," in Proceeding of Network and Distributed System Security Symposium (**NDSS**), 2020.

[C13] **Yun-Yun Tsai**, Pin-Yu Chen, Tsung-Yi Ho, "Adversarial Machine Learning for Social Good: Reprogramming black box machine learning model with scarce data and limited resources," Advances in Neural Information Processing Systems (**NeurIPS**) NewInML Workshop, Poster, 2019.

[C14] Ta-Wei Huang, **Yun-Yun Tsai**, Chung-Wei Lin, Tsung-Yi Ho, "Vehicle Sequence Reordering with Cooperative Adaptive Cruise

Control," in Proceeding of Design, Automation and Test in Europe Conference and Exhibition (**DATE**), 2019.

Patents

[C1] Pin-Yu Chen, **Yun-Yun Tsai**, Sijia Liu, Chia-Yu Chen, I-Hsin Chung, Tsung-Yi Ho. "Transfer Learning With Machine Learning Systems", U.S. Patent Application No: 17/029506, Application Date: September 23, 2020.

[C2] Pin-Yu Chen, I-Hsin Chung, Bo Wu, Chuang Gan, Lei Hsiung, **Yun-Yun Tsai**, Tsung-Yi Ho. "Composite Adversarial Attack Model Training For Neural Networks", U.S. Patent Application No: 18/331,211, Application Date: May 12, 2026.

HONORS AND AWARDS

- Student Travel Grant, CVPR, 2024.
- Scholar Award, NeurIPS, 2023.
- Student Travel Grant, CVPR, 2023.
- Ph.D. Dean's Fellowship, Columbia University, 2021.
- Best Presenter, Blackhat Award Forum, CyberSec Taiwan, 2020.

TEACHING EXPERIENCE

Teaching Assistant, Computer Vision II (COMS W4732) *Spring 2025*
Columbia University
• Instructor: Prof. Carl Vondrick
• Supported hands-on deep learning assignments, including implementing GANs and diffusion models from scratch.
• Guided students through model implementation, training, debugging, and evaluation.

Teaching Assistant, Engineering Software-as-a-Service (COMS W4152) *Fall 2023*
Columbia University
• Instructor: Prof. Junfeng Yang
• Supported programming assignments and team-based SaaS projects covering Ruby on Rails, agile development, and software testing (BDD/TDD).
• Assisted students through office hours, technical guidance, assignment grading, and project feedback.

⋆ *Courses Prepared to Teach:*
- **Undergraduate Courses:** Machine Learning, Computer Vision, Deep Learning, Software Engineering.
- **Graduate Courses:** Trustworthy and Secure AI, Unsupervised Learning, Generative AI.

INVITED TALKS AND CONFERENCE PRESENTATIONS

Invited Guest Lecture *December 7, 2025*
- AI Governance Course, Carnegie Mellon University, Silicon Valley
- Host: Prof. Lake Dai
- Talk Topic: AI Safety and Robustness of Modern AI Systems

Invited Guest Lecture *November 13, 2025*
- Frontiers of Tough Tech (IEME E4505), Columbia University
- Host: Prof. Harry West
- Talk Topic: Detecting AI-Generated Videos through Visual Artifacts and Multimodal Reasoning (REVEAL)

Invited Research Talk *January 7, 2025*
- Department of Electrical Engineering, National Taiwan University
- Host: Prof. Shao-Hua Sun
- Talk Topic: Exploring Safety of Machine Learning Models and Generative AI

Invited Conference Talk *December 7, 2024*
- Young Women Star Session, Taiwanese Association for Artificial Intelligence (TAAI 2024)
- Host: TAAI 2024 Organizing Committee
- Talk Topic: Research Journey in Trustworthy and Secure AI

Conference Presentation *December 5, 2020*

- Taiwanese Association for Artificial Intelligence (TAAI 2020)
- Host: TAAI 2020 Organizing Committee
- Talk Topic: Transfer Learning without Knowing: Reprogramming Black-Box Machine Learning Models

Invited Conference Talk *August 12, 2020*
- Blackhat Award Forum, CyberSec 2020, Taipei, Taiwan
- Host: CyberSec 2020 Blackhat Award Forum
- Talk Topic: CloudLeak: DNN Model Extraction from Commercial MLaaS Platforms

OUTREACH ACTIVITIES

Academic Service

Conference Reviewer: ICLR (2027); NeurIPS (2023–2025); ICML (2022–2025); ECCV (2024); ICCV (2024, 2025); CVPR (2023–2025); AAAI (2025–2027).

Research Mentoring

Mentored over 20 undergraduate and master's students in AI safety, computer vision, and machine learning, providing guidance on research methodology, experimental design, and academic writing.

Selected Mentees:

- Lei Hsiung (master), pursuing Ph.D. at Dartmouth College.
- Ruize Xu (master), pursuing Ph.D. at Dartmouth College.
- Qingyuan Liu (master), pursuing Ph.D. at Nanyang Technological University, Singapore.
- Jincen Song (master), AI Researcher at UC Berkeley.
- Jia Hao Liang (undergrad.), Undergraduate Student, Massachusetts Institute of Technology (MIT).
- Victoria Li (master), AI Engineer at Arcade AI
- Ruijian Zha (master), Software Engineer at Utopai Studios.
- Ju-chin (Peter) Zhao (master), Machine Learning Engineer at TikTok

Music Experience

Violist *2021-2022*
Columbia University Orchestra, New York, NY
- Performed at Lincoln Center's Alice Tully Hall in New York City.

First-Stand Violist *2014–2018*
National Tsing Hua University Symphony Orchestra, Hsinchu, Taiwan
- Served as first-stand violist for four years; performed in over 10 public concerts, including performances at Taiwan's National Concert Hall.

Piano Concerto Soloist *2017*
National Tsing Hua University, Hsinchu, Taiwan
- Winner of the university's Concerto Competition; performed Rachmaninoff's Piano Concerto No. 2 with the university orchestra at its annual summer concert.

ACADEMIC REFERENCES

Junfeng Yang (Ph.D. advisor)
Professor, ACM Fellow
Department of Computer Science
Columbia University
junfeng@cs.columbia.edu

Laurens van der Maaten (Mentor in Meta)
Member of Technical Staff
Anthropic
Research Scientist at Meta AI (Previously)
lvdmaaten@gmail.com

Chengzhi Mao (Research Collaborator)
Assistant Professor
Department of Computer Science
Rutgers University
chengzhi.mao@rutgers.edu

Tsung-Yi Ho (Master's advisor)
Professor, IEEE Fellow
Department of Computer Science and Engineering
The Chinese University of Hong Kong
tyho@cse.cuhk.edu.hk

Penghui Li (Research Collaborator)
Assistant Professor
College of Computing and Data Science
Nanyang Technological University, Singapore
penghui.li@ntu.edu.sg

Updated 2026-11-20.