Peiyang Xu

304B, Zijing Student Apartment Building 2, Tsinghua University, Beijing City, 100084, P.R.China
(+86)188-5256-9598 | xupy21@mails.tsinghua.edu.cn

Education

Department of Computer Science and Technology, Tsinghua University Beijing, China
Bachelor of Engineering in Computer Science and Technology Sep 2021 – Jun 2025

- GPA: 3.86/4.00 | Ranked 15th in the National College Entrance Examination (Jiangsu Province, top 0.01%)
- Courses: Software Engineering (A+), Introduction to Complex Analysis (A+), Student Research Training (A+), Modern Cryptography (A+), Computer Graphics (A), Artificial Neural Network (A)
- Research interest: Trustworthy Machine Learning, AI Safety, Multi-Modal Models

Research Experience

Project: SafeVision: Efficient Image Guardrail with Robust Policy Adherence and Explainability Chicago, IL
Research Assistant; Advisor: Bo Li, Associate Professor, Dept of CS, University of Chicago July – Nov 2024

- Proposed SafeVision, a novel image guardrail system that integrates human-like understanding and reasoning, with robust policy adherence, explainability, and lightning-fast inference speeds
- Provided VisionHARM-500K, a high-quality unsafe image benchmark comprising over 500k images to cover a wide array of risky categories, significantly broadening the scope and depth of unsafe image benchmarks
- Implemented an effective data collection and generation framework, a policy-following training pipeline, a customized loss function, and an efficient diverse QA generation strategy to enhance the training effectiveness
- Achieved state-of-the-art performance in both efficiency and accuracy, with an accuracy of 91.36% on VisionHARM-500K (17.36% higher than GPT-4O) and an inference time of 0.313 seconds per image
- Contributed to a first-author paper submitted to ICLR 2025

Project: MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models Beijing, China
Research Assistant; Advisor: Bo Li, Associate Professor, Dept of CS, University of Chicago Feb – Jun 2024

- Aimed to provide the first comprehensive safety and trustworthiness evaluation dataset and platform for multimodal foundation models (MMFMs), to help develop safer and more reliable MMFMs and systems
- Designed various evaluation scenarios/tasks and red teaming (attack) algorithms to generate challenging data and form a high-quality benchmark, to assess models from multiple perspectives, including safety, hallucination, fairness/bias, privacy, adversarial robustness, and out-of-distribution (OOD)
- Evaluated a range of state-of-the-art multimodal models, such as GPT-4o, Gemini, and Qwen, using our MMDT platform, and revealed several vulnerabilities and areas for improvement across all perspectives for these models
- Contributed to a paper submitted to ICLR 2025

Project: Natural Language Induced Adversarial Images Beijing, China
Research Assistant; Advisor: Xiaolin Hu, Associate Professor, Dept of CS, Tsinghua University July 2023 – April 2024

- Aimed to use prompt optimization to perform black-box attacks on commercial text-to-image models such as Midjourney and DALLE-3, by generating natural adversarial examples in order to let classifiers make mistakes
- Utilized and improved non-differentiable optimization methods such as Genetic Algorithm and Ant Colony Algorithm to optimize the input prompt in order to achieve the attacking purpose
- Performed Genetic Algorithm in the discrete prompt space and assessed the fitness of individuals using a weighted result of attack success rate and image naturalness, with the naturalness measured by CLIP
- First to explore using prompt-based optimization strategies for generating adversarial images, with an attack success rate of 92.1% and superior in terms of naturalness, transitivity, and interpretability
- Analyzed adversarial images from a natural language perspective and demonstrated that the adversarial semantic information extracted from these images significantly affects the accuracy of traditional image classifiers
- Contributed to a co-first author paper accepted by the 2024 ACM Multimedia (ACM MM 2024)

Project: CurBench: Curriculum Learning Benchmark Beijing, China
Research Assistant; Advisor: Xin Wang, Assistant Researcher, Dept of CS, Tsinghua University Oct 2022 – Jan 2023

- Aimed to use curriculum learning to improve various classification algorithms and to establish the first benchmark for a systematic evaluation on curriculum learning
- Created a benchmark consisting of 16 datasets spanning 3 research domains (computer vision, natural language processing, and graph machine learning) under 3 settings (standard, noised, and imbalanced)
- Provided a unified pipeline that plugs automatic curricula into general machine learning process, supporting 14 core curriculum learning methods (personally responsible for the Meta Weight Net)
- Contributed to a paper accepted by the 2024 ICML conference

Publications & Manuscripts

- **Peiyang Xu**, Minzhou Pan, Zhaorun Chen, Xue Lin, Chaowei Xiao, Bo Li, SafeVision: Efficient Image Guardrail with Robust Policy Adherence and Explainability (submitted to ICLR 2025)
- Chejian Xu, Jiawei Zhang , Zhaorun Chen, Chulin Xie, Mintong Kang, Zhuowen Yuan, Zidi Xiong, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, **Peiyang Xu**, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Zhen Xiang, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li, MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models (submitted to ICLR 2025)

- Xiaopei Zhu*, **Peiyang Xu***, Guanning Zeng, Yinpeng Dong, Xiaolin Hu, Natural Language Induced Adversarial Images (ACM MM 2024).
- Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, **Peiyang Xu**, Shengnan Liu and Wenwu Zhu, CurBench: Curriculum Learning Benchmark, accepted by 42th International Conference on Machine Learning (ICML 2024)

Honors & Awards

- Tsinghua University Research Excellence Scholarship (2024)
- Tsinghua University Academic Excellence Scholarship (2024)

Skills & Others

Language: Chinese (Native); English (Fluent)
Exams: TOFEL: 108 (Speaking 23); GRE: 158+170+4.0
Computer Science: Proficient in Python, C++, Rust, and various other programming languages; Skilled in using deep learning libraries like PyTorch; Deep understanding of multimodal models