Shunchang Liu

Rte Cantonale, 1015 Lausanne, Switzerland
shunchangliu@outlook.com — (+86)18810220881 — Google Scholar — https://shunchang-liu.github.io

RESEARCH INTERESTS

I'm interested in studying **Safety & Privacy** in complex AI models (e.g., LLMs/LVLMs), focusing on challenges such as adversarial attack, reward hacking, and copyright issues. I am always driven by the long-term goal of building *explainable* and *provable* safe AI systems.

EDUCATION

EPF Lausanne - ETH Zurich (joint-degree program), Lausanne & Zurich, Switzerland — Sep 2024 — Present
Master of Computer Science: Cybersecurity — GPA: 5.63 / 6.00

Beihang University, Beijing, China — Sep 2021 — Dec 2023
Graduate Research Assistant

Beihang University, Beijing, China — Sep 2017 — Jun 2021
Bachelor of Engineering: Automation — GPA: 3.87 / 4.00

WORK EXPERIENCES

Concordia AI (Social Enterprise focused on AI Safety and Governance) — Beijing, China
Intern, Mentor: Brian Tse — Apr 2024 — Aug 2024, Jun 2025 — Aug 2025

- Contributed to two reports, [Frontier AI Risk Management Framework and [Responsible Innovation in AI × Life Sciences, which were launched at WAIC 2025
- Implemented a frontier AI risk assessment platform based on Inspect Evals, supporting automated evaluations of misuse risks, model honesty, and self-awareness [Project Page]

PUBLICATIONS

Co-first author (*) / Corresponding author (†)

The Biased Oracle: Assessing LLMs' Understandability and Empathy in Medical Diagnoses. Jianzhou Yao*, **Shunchang Liu***, Guillaume Drui, Rikard Pettersson, Alessandro Blasimme, Sara Kijewski. *NeurIPS Workshop @GenAI4Health*, 2025 [paper] [code]

CopyJudge: Automated Copyright Infringement Identification and Mitigation in Text-to-Image Diffusion Models. **Shunchang Liu***, Zhuan Shi*, Lingjuan Lyu, Yaochu Jin, Boi Faltings. *ACM Multimedia*, 2025 [paper] [code]

Boosting Cross-task Transferability of Adversarial Patches with Visual Relations. Tony Ma, Songze Li, Yisong Xiao, **Shunchang Liu†**. *IEEE CVPR Workshop @AdvCV*, 2023 [paper]

Benchmarking the Robustness of Quantized Models. Yisong Xiao, Tianyuan Zhang, **Shunchang Liu**, Haotong Qin. *IEEE CVPR Workshop @AdvCV*, 2023 [paper]

Harnessing Perceptual Adversarial Patches for Crowd Counting. **Shunchang Liu***, Jiakai Wang*, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, Dacheng Tao. *ACM CCS*, 2022 [paper] [code]

Revisiting Audio Visual Scene-Aware Dialog. Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, **Shunchang Liu**. *Neurocomputing*, 2022 [paper]

Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World. Jiakai Wang, Aishan Liu, Zixin Yin, **Shunchang Liu**, Shiyu Tang, Xianglong Liu. *IEEE CVPR*, 2021 (**Oral**) [paper] [code]

RESEARCH EXPERIENCES

Reward Hacking Mitigation in RLHF via Sparse Autoencoder ETH Zurich
Semester Project, Supervisor: Prof. Andreas Krause and Dr. Xin Chen Mar 2025 — Present

- Used sparse autoencoder (SAE) to construct an interpretable preference model and employing causal intervention on SAE features to mitigate reward hacking during RLHF (in progress)

Copyright Infringement Detection and Mitigation for Diffusion Models EPFL
Semester Project, Supervisor: Prof. Boi Faltings and Dr. Zhuan Shi Sep 2024 — Jan 2025

- Proposed CopyJudge, an automated framework that uses large vision-language models to detect copyright infringement by comparing AI-generated and copyrighted images using abstraction-filtration-comparison and multi-model debate, and reducing risk through prompt tuning and non-infringing noise exploration in latent space.
- Published one [paper] on ACM MM

Cross-task Adversarial Patch Generation Beihang University
Research Assistant, Supervisor: Dr. Jiakai Wang Jan 2023 — Jun 2023

- Proposed a novel visual relation-based adversarial patch generation method combing object misclassification and predicate-based relation elimination, which improved black-box adversarial transferability across diverse visual reasoning tasks such as image captioning and visual question answering
- Published one [paper] on IEEE CVPR Workshop

Adversarial Patch Generation for Crowd Counting Beihang University
Research Assistant, Supervisor: Prof. Aishan Liu and Dr. Yingwei Li (JHU) Jun 2021 — Aug 2022

- Proposed a novel perceptual adversarial patch generation framework that exploited model-inherent perceptual properties, e.g., scale and position perceptions, of crowd counting models, which led to the SOTA attack
- Employed adversarial training with our patches to improve models' cross-dataset generalization and robustness towards complex backgrounds, which showcased evidence of its beneficial impact on vanilla models' performance
- Published one [paper] on ACM CCS

Adversarial Camouflage Generation in Physical World Beihang University
Research Assistant, Supervisor: Dr. Jiakai Wang July 2020 — Nov 2020

- Proposed a novel dual attention suppression attack to generate visually natural adversarial vehicle camouflages by evading both model-shared attention and human-specific attention, which achieved state-of-the-art black-box attacking performance in both digital and physical world towards classification and detection tasks
- Published one [paper] on IEEE CVPR

AWARDS

- Outstanding Graduate of Beihang University **(Top 20%)** Jun 2021
- CVPR Security AI Challenge: No-limit Adversarial Attacks on ImageNet **(11/1559)** Apr 2021
- First Prize of Discipline Competition Scholarship, BUAA Dec 2019
- First Prize of Study Excellence Scholarship, BUAA Dec 2019
- Outstanding Student of Beihang University **(Top 5%)** Jun 2019
- Second Prize of Group A of Non-physics in the 35th National Physics Competition for College Students Dec 2018

SERVICES

- **Program Committee / Reviewer:** Pattern Recognition, Frontiers of Computer Science, IEEE TCSVT, IEEE T-ITS, and Workshops (ICML MoFA 2025, CVPR AdvCV 2023, etc.)
- **Teaching Assistant:** Machine Learning (EPFL Fall 2025)

SKILLS

- **Programming:** C, Python, R, etc.
- **Software:** Matlab, Pytorch, TensorFlow, MindSpore, etc.
- **Language:** Chinese, English (IELTS Academic overall 7.0, listening 6.5, reading 8.0, speaking 6.0, writing 7.0)