杨天韵

tianyunyoung@hotmail.com 188 0107 0774 Google Scholar
6 Kexueyuan South Road, Haidian, Beijing, China WeChat: yty15071227330 Website

Education

Institute of Computing Technology, Chinese Academy of Sciences *Sep. 2019 - June 2025 (expected)*
Ph.D., School of Computer Science.

- GPA: 3.5/4.0
- **Advisor:** Juan Cao.

The University of Sydney *Sep. 2023 - Sep. 2024*
Joint Ph.D, School of Computer Science.

- **Advisor:** Chang Xu.

Wuhan University *Sep. 2015 - June 2019*
B.E., School of Electrical Engineering, Excellent Engineer Class

- GPA: 3.81/4, Top 5%

Research Interests

My research focuses on AGI safety and mechanistic interpretability, covering topics such as hallucination mitigation, concept editing, and model attribution:

- **Hallucination Mitigation**: My work adopts a modular perspective to investigate the causes of hallucination in large vision-language models, analyzing how their components contribute to this issue and proposing training-free and training-based methods to mitigate it (ICLR 2025).
- **Concept Editing**: My work designs a robust concept erasing method based on differential pruning to eliminate harmful or copyrighted concepts from diffusion models (NeurIPS 2024 Safe Generative AI Workshop).
- **Model Attribution**: My research aims to identify the source generative model of AI-generated content by extracting the unique "fingerprints" left by the model during the generation process. This includes work on model architecture attribution (AAAI 2022), open-set model attribution (CVPR 2023), and zero-shot model attribution (TMM 2025).

Publications

- **Tianyun Yang**, Ziniu Li, Juan Cao, Chang Xu. *Mitigating Hallucinations in Large-Vision Language Models via Modular Attribution and Intervention.* International Conference on Learning Representations (ICLR), 2025.
- **Tianyun Yang**, Juan Cao, Danding Wang, Chang Xu. *Model Synthesis for Zero-shot Model Attribution.* IEEE Transactions on Multimedia (TMM), 2025.
- **Tianyun Yang**, Ziniu Li, Juan Cao, Chang Xu. *Pruning for Robust Concept Erasing in Diffusion Models.* Workshop on Safe Generative AI at Conference on Neural Information Processing System (NeurIPS), 2024.
- **Tianyun Yang**, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, Sheng Tang. *Progressive Open Space Expansion for Open-Set Model Attribution.* The IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR) 2023.

- **Tianyun Yang**, Ziyao Huang, Juan Cao, Lei Li, Xirong Li. *Deepfake Network Architecture Attribution.* Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI) 2022.
- Peng Qi, Juan Cao, **Tianyun Yang**, Junbo Guo, Jintao Li. *Exploiting Multi-Domain Visual Information for Fake News Detection.* IEEE International Conference on Data Mining (ICDM) 2019.
- Juan Cao, Peng Qi, Qiang Sheng, **Tianyun Yang**, Junbo Guo, Jintao Li. *Exploring the Role of Visual Content in Fake News Detection.* Lecture Notes in Social Networks, 2020

Manuscripts

- **Tianyun Yang**, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, Sheng Tang *Learning to Disentangle GAN Fingerprint for Fake Image Attribution* arxiv:2106.08749.

Service

- Reviewer: T-MM, ICLR'25, TMLR'24, NeurIPS'24, NeurIPS'23, ICLR'24, CVPR'23, NeurIPS'22
- Teaching Assistant: Multimedia Technology, UCAS, Spring 2022

Awards

- First Prize of Academic Award, University of Chinese Academy of Sciences 2022
- Director's Excellence Scholarship, Institute of Computing Technology 2021
- The 1st Prize in Chinese AI Competition, Deepfake Identification 2021
- The 1st Prize in China Undergraduate Mathematical Contest in Modeling, Hubei Province 2018
- Outstanding Student, Wuhan University 2017