Yu Huang

+86 13330936878 | yhuang489@connect.hkust-gz.edu.cn | homepage | hardenyu21

Education Background

- **Hong Kong University of Science and Technology (Guangzhou)** *09. 2024 - 07. 2026*
  ***MPhil** in Artificial Intelligence | **GPA**: 3.65 / 4.3* Guangzhou, China
  ◦ Supervisor: Prof. Xuming Hu, Research Interest: Trustworthy AI, MLLM

- **The University of Hong Kong** *09. 2022 - 07. 2024*
  ***MS** in Data Science | **GPA**: 3.51 / 4.3* Hong Kong
  ◦ Thesis Project: A benchmarking on deep learning based medical report generation

- **Chongqing University** *09. 2018 - 07. 2022*
  ***B.Eeng** in Electronic Information Engineering | **GPA**: 86.26 / 100* Chongqing, China

Research Experience

- **HKUST-GZ NLP Group** *09.2024 - 07. 2026*
  Research Postgraduate | Trustworthy AI Guangzhou, China

  Conduct research on **trustworthy AI**, mainly focus on (i) developing faithful and challenging benchmarks for LLM evaluation and (ii) designing watermarking methods for AIGC, to address the broader challenge of reliable and secure deployment of advanced AI models in real-world scenarios.

  ◦ Develop **KnowMT-Bench**, the first-ever benchmark to systematically evaluate LLMs in **multi-turn** long-form question answering across **knowledge-intensive** fields

  ◦ Propose **Video Signature** (VIDSIG), an **in-generation** watermarking method that can embed robust and imperceptible watermark into the generated videos. VIDSIG achieves the best overall performance in watermark extraction accuracy, robustness, video quality and efficiency compared to other methods

Publications

*: Equal Contribution

- KnowMT-Bench: Benchmarking Knowledge-Intensive Long-Form Question Answering in Multi-Turn Dialogues
  Junhao Chen, **Yu Huang***, Siyuan LI, Hanqian Li, Rui Yao, Hanyu Zhang, Jungang Li, Jian Chen, Bowen Wang, Xuming Hu
  [ICLR'2026 Under Review] [paper] [code]
- Video Signature: In-generation Watermarking for Latent Video Diffusion Models
  Yu Huang, Junhao Chen, Shuliang Liu, Hanqian Li, Qi Zheng, Yi R. Fung, Xuming Hu
  [AAAI'2026 Under Review, Phase I passed] [paper] [code]
- Videomark: A distortion-free robust watermarking framework for video diffusion models
  Xuming Hu, Hanqian Li, Jungang Li, **Yu Huang**, Aiwei Liu
  [AAAI'2026 Under Review, Phase I passed] [paper] [code]
- Guarding the gate: Conceptguard battles concept-level backdoors in concept bottleneck models
  Songning Lai, **Yu Huang***, Jiayu Yang, Gaoxiang Huang, Wenshuo Chen, Yutao Yue
  [ICLR'2026 Under Review] [paper]
- Multimodal Deception in Explainable AI: Concept-Level Backdoor Attacks on Concept Bottleneck Models
  Songning Lai, Jiayu Yang, **Yu Huang***, Lijie Hu, Tianlang Xue, Zhangyi Hu, Jiaxu Li, Haicheng Liao, Yutao Yue
  [TMLR Under Review] [paper]

Honors and Awards

- Red Bird Post-Graduate Scholarship, HKUST(GZ) 2024
- Outstanding Student Scholarship, Chongqing University 2021
- Outstanding Student Award, Chongqing Univeristy 2021
- Second-class Scholarship, Chongqing Univeristy 2020

Academic Activities

- Reviewer for AAAI'2026
- Reviewer for AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)