黄启栋

手机：13085060686 | 邮箱：hqd0037@mail.ustc.edu.cn
微信：15305851516 | 个人主页：https://shikiw.github.io/

教育经历

中国科学技术大学 985 双一流 2020年09月 - 2025年06月

网络空间安全 博士 网络空间安全学院 合肥

GPA：3.49/4.3 成绩排名前20%

中国科学技术大学 985 双一流 2016年09月 - 2020年06月

信息安全 本科 信息科学技术学院 合肥

GPA：3.31/4.3 成绩排名前50%

实习经历

多模态大模型基础研究 2023年08月 - 至今

研究型实习生——上海人工智能实验室 上海

主要围绕多模态大模型基础模型方面展开研究，包括但不限于多模态幻觉、模型训练/推理加速、模型训练指标评估设计、跨模态对齐等，有一定多模态大模型多机多卡训练调优经验。在职期间已发表CVPR 2024一篇并被入选Highlight论文，在投三篇。

基于条件引导DIffusion的中文文生图模型训练研究 2022年05月 - 2022年07月

研究型实习生——科大讯飞研究院 合肥

主要围绕条件引导DIffusion在大规模图文数据集（如Wukong）上进行训练。

研究兴趣及技能

多模态大模型

主要集中在多模态大语言模型方面，包括但不限于多模态幻觉、模型训练/推理加速、模型训练指标评估设计等。已作为第一作者，已在CCF A类计算机视觉会议CVPR 2024上发表了一篇论文，即OPERA。该研究深入探讨了多模态LLMs产生幻觉的深层原因，并提出了信息衰减的解释。在此基础上，我们提出了“过度信赖注意力惩罚”和“回溯-分配机制”，以减轻多模态幻觉问题。该工作在社交平台阅读量有超过5万阅读量与4千转发数，半年引用超过60。此外，主导了一项大型视觉语言模型预训练评估指标设计工作MIR，一项大型视觉语言模型跨模态对齐工作MoCa,一项基于视觉冗余的大型视觉语言模型训练/推理加速工作PyramidDrop，以及一项大语言模型越狱检测的工作RDPI，具备一定的多机多卡训练调优经验。

可信高效AI

在可信AI领域我主要围绕对抗攻击、后门攻击、大模型越狱攻击、以及AIGC生成内容安全等方面展开工作。我已在CCF A类计算机视觉会议CVPR、ICCV、AAAI等发表了四篇第一作者论文和若干篇合作论文。在对抗样本领域我有以下工作：1）针对2D图像无监督预训练模型的对抗防御方法RobustMAE；2）针对3D模型的对抗攻击/防御方法SI-Adv/PointCAT。在后门攻击领域我有针对2D图像的不可见后门攻击方法Poison Ink。在大模型越狱攻击领域有在投的越狱检测防御工作一项。在AIGC内容安全领域有针对深度换脸的“主动防御”工作一项，以及针对基于Stable Diffusion的文生图模型的图像内容安全工作SimAC。（其中，我们是首个提出“主动防御”DeepFake的概念的团队，通过在操纵前主动保护用户的面部隐私，不同于之前的如DeepFake检测等事后应对措施）。

在高效AI方面，我已发表一篇CCF A类计算机视觉顶会CVPR 2023。该工作专注于研究大规模视觉预训练模型，提出了DAM-VP。该方法通过一种基于多样性感知的高效且自适应的视觉提示学习方法，解决了视觉提示与下游数据多样性之间的不匹配问题。

研究成果

第一作者论文

1. Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu. Deciphering Cross-Modal Alignment in Large Vision-Language Models with Modality Integration Rate. **(Under Review at a top-tier conference)**

2. Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu. OPERA : Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024. **(Selected as Highlight Paper, 2.8% of submissions)**

3. Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, Kui Zhang, Gang Hua, Nenghai Yu. PointCAT : Contrastive Adversarial Training for Robust Point Cloud Recognition. IEEE Transactions on Image Processing (**TIP**), 2024.

4. Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, Nenghai Yu. Improving Adversarial Robustness of Masked Autoencoders via Test-time Frequency-domain Prompting. IEEE/CVF International Conference on Computer Vision (**ICCV**), 2023.

5. Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, Nenghai Yu. Diversity-Aware Meta Visual Prompting. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2023.

6. Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, Nenghai Yu. Shapeinvariant 3D Adversarial Point Clouds. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022.

7. Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, Nenghai Yu. Initiative Defense against Facial Manipulation. AAAI Conference on Artificial Intelligence (**AAAI**), 2021.

合作论文

1. Long Xing, **Qidong Huang**, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, Dahua Lin. PyramidDrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. **(Under Review)**

2. Likai Liang, **Qidong Huang**, Weiming Zhang, Wenying Zhang. RDPI: Defending against Multi-Turn Jailbreak Attacks via Response-Based Dynamic Prompt Inference. **(Under Review)**

3. Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, **Qidong Huang†** . SimAC : A Simple AntiCustomization Method against Text-to-Image Synthesis of Diffusion Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024. **(† Corresponding author)**

4. Kui Zhang, Hang Zhou, Jie Zhang, **Qidong Huang**, Weiming Zhang, Nenghai Yu. Ada3Diff : Defending against 3D Adversarial Point Clouds via Adaptive Diffusion. ACM International Conference on Multimedia (**MM**), 2023

5. Han Fang, Dongdong Chen, **Qidong Huang**, Jie Zhang, Zehua Ma, Weiming Zhang and Nenghai Yu. Deep Template-based Watermarking. IEEE Transactions on Circuits and Systems for Video Technology (**TCSVT**), 2020.

6. Jie Zhang, Dongdong Chen, **Qidong Huang**, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, Nenghai Yu. Poison ink : Robust and invisible backdoor attack. IEEE Transactions on Image Processing (**TIP**), 2022.

荣誉奖项

国家奖学金	2021年
安恒信息奖学金	2023年
“互联网+”安徽省大学生创新创业大赛省铜奖	2023年
校优秀共青团员	2021年
中国科学技术大学研究生学业奖学金	2020-2024年
全国中学生数学联赛省一等奖	2014年，2015年

技能/证书及其他

- **技能：** 数理基础扎实，掌握Python等语言，熟练使用Pytorch框架，有较好的深度学习理论基础
- **语言：** 普通话（母语），英语（CET-6）

其他经历

校人工智能俱乐部 2020年07月 - 2022年06月

担任副主席，协助开展相关活动（专题讲座、交流论坛、论文分享会等），给全校同学带来更多人工智能领域的专题分享。

受邀讲座 (AI SPOT, OpenMMLab)

分享自己的CVPR工作，主题为Exploring MLLM’s Hallucination from A Causal Attention Perspective。多平台同步直播，观看人数总计破万。

顶级会议/期刊审稿人

担任包括TPAMI、TIP、TNNLS、CVPR、ICCV、ECCV、ICLR、NeurIPS在内人工智能方面国际顶级会议期刊的审稿人。