香港城市大學
City University of Hong Kong

游珅

15123006339 |1391419813@qq.com|男

教育经历

- **香港城市大学 博士研究生(全日制全额奖学金)** 计算机与科学 计算机与科学学院 **2024.09 —— 2028.06**
  GPA:3.85/4.3
- **电子科技大学 硕士研究生** 软件工程 信息与软件工程学院 **2021.09 —— 2024.06**
  GPA:3.51/4.0 **综合排名: 1/232**
- **西南交通大学 本科** 主修: 生命科学与工程学院制药工程(化学制药) 辅修: 软件工程 **2017.09 —— 2021.06**
  GPA:3.12/4.0

实习经历

- **阿里巴巴-淘天集团-淘宝直播-直播客户端技术** **2023.06 —— 2024.03**
  - 负责直播链路中的操作打点和性能监控统计，完善了操作打点的缺失部分，并添加了程序异常和性能异常上报功能。
  - 使用C++、FFMpeg、OpenGL、GDI32开发高性能特效播放器(AlphaVideo)，将解码、像素格式转换、图像通道处理、缩放等操作迁移至GPU执行。相较基于WebGL实现的播放器，该特效播放器CPU消耗降低**90%**，内存占用节约**50%**。
- **香港理工大学-科研助理** **2023.06 —— 2024.12**
  - 负责调研并设计图神经网络的后门攻击方法，主要针对GCN、GAT等自生成图嵌入模型进行攻击。基于图卷积特性，使用遗传算法和梯度下降法分别对节点和边进行扰动，同时利用节点梯度信息和度信息指导对抗样本生成。最终算法生成时间仅为当前最优算法的**12.5%**，攻击成功率超越现有最优算法**30%**

项目经历

- **大语言模型对抗性攻击** **2025.1 —— 2025.5**
  针对大语言模型在决策边界覆盖不全导致易受对抗攻击的安全漏洞问题，提出了一种统一的token级越狱攻击方法来揭示模型脆弱性。设计了基于进化算法的扰动集优化框架，结合梯度和交叉算子提升多场景攻击性能，并开发了扰动复用机制和token敏感性分析实现零样本攻击。在三种不同模型上的基准实验显示，相比基线方法攻击成功率分别提升62.36%、57.89%和64.81%。
- **医学大语言模型的安全性评估** **2025.6 —— 2025.9**
  针对医疗领域大语言模型在非语义输入变化下鲁棒性评估不足的问题，提出了基于对抗攻击的系统性评估框架。采用字符级替换策略模拟真实威胁场景（包括用户错误和恶意攻击），并建立了符合医疗LLM操作指南的分层评估指标体系，实现跨严重等级的多维安全评估。在多种医疗LLM上的实验表明，相比现有方法该攻击算法表现更优，能够准确评估各类医疗LLM的鲁棒性，为开发可靠的医疗AI系统提供支撑。
- **大语言模型针对性防御** **2025.10 —— 2026.1**
  针对大语言模型容易受到越狱攻击绕过安全机制且现有防御方法计算开销大或损害功能的问题，开发了高效的参数干预防护框架。通过梯度追踪技术精确定位安全关键神经元，仅修改上千个参数即可显著提升安全性。实验结果显示攻击成功率相比基线降低 77.1%，相比现有最优方法降低 22%-43%，同时保持 99%原始功能性能。

荣誉奖项

- **硕士研究生国家奖学金**(<4%) 2023.11
- **香港城市大学研究奖学金**(<10%) 2025.9
- **四川省优秀研究生毕业生**(<4%) 2023.11
- **电子科技大学优秀研究生毕业生**(<10%) 2023.11
- **电子科技大学硕士研究生奖学金** 一等奖(<7%) 2023.11
- **电子科技大学硕士研究生奖学金** 一等奖(<7%) 2022.11
- **电子科技大学硕士研究生奖学金** 二等奖(<30%) 2021.11

● **第十二届全国大学生高教社杯数学建模**	一等奖(<8%)	2019.10
● **第十三届全国大学生高教社杯数学建模**	一等奖(<8%)	2020.10
● **第十六届五一数学建模竞赛**	一等奖(<5%)	2019.06
● 第十八届全国研究生数学建模竞赛	三等奖	2021.10
● 第十九届全国研究生数学建模竞赛	三等奖	2022.10
● 中国高校计算机大赛-团体程序设计天梯赛	二等奖	2018.04

科研产出

●**接收**: Query-Efficient Generation of Adversarial Examples for Defensive DNNs via Multi-Objective Optimization(**学生一作, TEVC, SCI一区**)

●**接收**: SOPA: Sensitivity-Oriented Poisoning Attacks for Self-Supervised Graph Embedding Model via Bilevel Evolutionary Optimization(**第一作者, TEVC, SCI一区)**

●**接收**: UniBreak: A Unified Evolutionary Token-Level Jailbreaking Framework for Large Language Models (**第一作者,TEVC special issue, SCI一区)**

●**接收**: Iterative Training Attack : A Black-Box Adversarial Attack via Perturbation Generative Network (学生二作, SCI 四区)

●**接收**: Elucidating Spatiotemporal Chromatin Dynamics with Multi-stage Differential Variations from Hi-C(第三作者,KBS,SCI一区)

●在投: PinPoint: Neural Parameter Editing for Precise Jailbreak Defenses in Large Language Models (**第一作者,TNNLS ,SCI一区**)

●在投: Physically Real-time Infrared Attack against Optical Flow Estimation Networks (**第一作者,DSN2027,CCFB**)

●在投: Towards Secure Healthcare AI Agents: A Robust Evaluation Framework for Large Language Models (**第一作者,JBHI,SCI一区**)

●在投: VEAttack: Downstream-agnostic Vision Encoder Attack against Large Vision Language Models (第三作者,ICLR2026,CCFA)

●在投: METRON: Metabolic Dynamic Perception Kolmogorov-Arnold Network for Biological Age Estimation (第三作者,TCBB,JCRQ1)

●在投: RiboMUSE: Unified Multi-view Sequence-Structure Representation Learning for Internal Ribosome Entry Site Annotation (第三作者,TNNLS,SCI一区)

●在投: VEAttack: Downstream-agnostic Vision Encoder Attack against Large Vision Language Models (第三作者,ICLR2026,CCFA)

研究方向

对抗机器学习：高效对抗样本生成、自监督模型投毒攻击及深度神经网络防御

AI安全：大语言模型越狱攻击防御、真实对抗扰动及安全评估框架

进化计算：多目标优化、双层优化及机器学习进化算法应用