Zihan Guan

Ph.D. Student
Department of Computer Science
The University of Virginia

University of Virginia
☎ *(+1) 919-358-2817*
✉ *bxv6gs@virginia.edu*
Webpage
Github *Linkedin* *Google Scholar*

Research Interests

My research interests lie in Large Foundation Models, Generative AI, and Epidemic Forecasting. I have previous experience in **training and fine-tuning** large models such as Diffusion Models (e.g., Stable-Diffusion), Segment-Anything-Model (SAM), and LLama over a large scale of datasets.

Education

2023–present **PhD, Computer Science**, *University of Virginia*, Charlottesville, VA.
Advisor: Anil Vullikanti, GPA: 3.90/4.00

2020–2021 : **Master of Science, Computing**, *Imperial College London*, London, UK.
Thesis: Scalable Methods for Neural Network Verification
Advisor: Yang Zheng and Alessio Lomuscio

2016–2020 : **Bachelor of Management, Logistics Management**, *Dalian University of Technology*, China.
GPA: **3.92/4.00**, Rank: **1/29**

Selected Projects

June,2024-Oct,2024 **Investigating Bias and Unfairness in RAG-based Large Language Models**, [PDF].
1. Investigated fairness risks in RAG-based LLMs using a novel three-level real-world threat model.
2. Developed RAG systems with dense retrieval and conducted extensive experiments demonstrating fairness degradation across various tasks, including classification, question answering, and generation, using Llama7B, Llama13B, GPT-4, and GPT-4-mini.
3. Explored components of RAG, such as pre-retrieval and post-retrieval strategies, as potential mitigation approaches and emphasized the urgent need for robust fairness safeguards in RAG-based LLMs.

Dec,2023-Feb,2024 **A Unified Framework for Black-box Backdoor Detection on Diffusion Models**, [PDF].
1. Developed a novel causal inference framework to analyze backdoor attacks on diffusion models.
2. Designed a perturbation-based black-box detection method applicable to both conditional (e.g., text-to-image Stable Diffusion) and unconditional diffusion models (e.g., DDPM).
3. Achieved $\geq 92\%$ AUROC in detection effectiveness with only limited runtime overhead.

May,2023-June,2023 **Investigating Security Vulnerabilities on Segment-Anything-Models (SAM)**, [PDF].
1. Developed BadSAM, the first backdoor attack on the SAM for image segmentation.
2. Injected backdoors into the vanilla SAM leveraging SAM-adapter under model customization scenario.
3. Validated BadSAM's effectiveness on the CAMO dataset, showing high performance on clean samples and strong attack efficacy on poisoned inputs. Relevant findings are published in AAAI '24.

Aug,2022-May,2023 **Securing Deep Neural Networks against Backdoor Attacks**, [PDF].
1. Investigated the early-fitting phenomenon in backdoor attacks and introduced a novel synchronization concept based on model explainability.
2. Developed the Deep Backdoor Attack (DBA) to evade existing defenses by encouraging deep-layer synchronization and suppressing shallow-layer synchronization.
3. Validated DBA's effectiveness and robustness across multiple datasets and tasks (e.g., CIFAR-10) on popular model architectures such as ResNet and EfficientNet. Relevant results are published in CIKM '23.

Work Experience

Aug,2023 – Now ***University of Virginia, Biocomplexity Institute***, *Research Assistant*, Charlottesville, VA.
Focus on privacy and safety of **GenAI**, **LLMs** and **Epidemics Forecasting**.
Advisor: Dr. Anil Vullikanti

Aug,2022 – May,2023 ***University of Georgia***, *Research Assistant*, Athens, GA.
Focus on backdoor defense and attacks in neural networks.
Host: Dr. Ninghao Liu

Dec,2019 – Mar,2020 ***Mobu***, *Android Engineer (Intern)*, Shanghai, China.
Developed more than 10 APPs independently, including Compass, Booster, and Mini Games; Participated in the development and release of the Apps in the overseas market.

Dec,2018 – Mar,2019 ***Accenture***, *Software Engineer (Intern)*, Dalian, China.
Engaged in BMW project in the automobile industry, developing an internal inventory system; Responsible for the layout design and business logistics of the log-in page; Received and fed back to the client's demand, completed 2 rounds of iterative development.

Selected Publications & Preprints

Safety in Diffusion Models [1], SAM [3], Graph Neural Networks [4], and Deep Neural Networks [5, 2, 6]; LLMs/RAG in Fairness [9], Healthcare [10] and Geography [8, 7].

Publications

[1] **Zihan Guan***, Mengxuan Hu*, Sheng Li, and Anil Vullikanti. Ufid: A unified framework for input-level backdoor detection on diffusion models. ***AAAI***, **2025**.

[2] Mengxuan Hu*, **Zihan Guan***, Zhongliang Zhou, Jielu Zhang, and Sheng Li. BBCal: Black-box backdoor detection under the causality lens. ***TMLR***, **2024**. Featured Certification 🏆.

[3] **Zihan Guan**, Mengxuan Hu, Zhongliang Zhou, Jielu Zhang, Sheng Li, and Ninghao Liu. Badsam: Exploring security vulnerabilities of sam via backdoor attacks. ***AAAI***, **2024**.

[4] **Zihan Guan**, Mengnan Du, and Ninghao Liu. Xgbd: Explanation-guided graph backdoor detection. ***ECAI***, **2023**.

[5] **Zihan Guan**, Lichao Sun, Mengnan Du, and Ninghao Liu. Attacking neural networks with neural networks: Towards deep synchronization for backdoor attacks. ***CIKM***, **2023**.

[6] Yucheng Shi, Mengnan Du, Xuansheng Wu, **Zihan Guan**, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. ***NeurIPS***, **2023**.

[7] Zhongliang Zhou, Jielu Zhang, **Zihan Guan**, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. ***SIGIR***, **2024**.

[8] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Mengxuan Hu, **Zihan Guan**, Sheng Li, and Lan Mu. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. ***ACM SIGSPATIAL GeoAI Workshop***, **2024**.

Under Reviews and Arxiv

[9] Mengxuan Hu, Hongyi Wu, **Zihan Guan**, Ronghang Zhu, Dongliang Guo, Daiqing Qi, and Sheng Li. No free lunch: Retrieval-augmented generation undermines fairness in llms, even for vigilant users. *Submitted to ICLR*, 2024. Under review.

[10] **Zihan Guan**, Zihao Wu, Zhengliang Liu, Dufan Wu, Hui Ren, Quanzheng Li, Xiang Li, and Ninghao Liu. Cohortgpt: An enhanced gpt for participant recruitment in clinical study, 2023.

Fellowships & Awards

2024 1st Winner of PETs for Public Heath Challenge (Link), **$50,000**

2023 Recipient of ***UVA Computer Science Scholar*** for the first year Ph.D. study.

2020 Excellent Student Awards, Dalian (top 5%)

2019 ***Meritorious Winner*** of Mathematical Contest In Modeling contest.

2017-2018 Technology and Information Scholarship , DUT (top 5%)

2017-2019 Academic Excellence Scholarship, DUT (GPA top 5%)

2017-2018 Merit Student, DUT

Professional Services

2024 The Thirteenth International Conference on Learning Representations (ICLR 2025), Reviewer

2024 The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Reviewer

2023 Intelligent Data Analysis, Reviewer

2023 The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Reviewer

2023 26th European Conference on Artificial Intelligence (ECAI), Reviewer

Teaching Experiences

2024 Data Structures and Algorithms (CS 2100), Undergraduate Level, Teaching Assistant