Bowei Tian

btian1@umd.edu ◇ Website ◇ LinkedIn ◇ College Park, MD, USA

PROFILE

Experienced in AI research and development, focusing on large language models, interpretability (causal, representation learning), and computer vision. Recognized through publications and hands-on research, skilled in **Python, C++, SQL,** and **HTML/CSS/JavaScript**. Focused on Representation Engineering in LLMs and trustworthy AI, demonstrating a commitment to building robust and interpretable AI systems.

EDUCATION

◆ University of Maryland, College Park, MD, USA

First-year Ph.D. student of Electrical and Computer Engineering, expected in May 2029

◆ Wuhan University, Wuhan, CHN

Bachelor's degree in Engineering in Information Security, received in June 2024

Cumulative GPA: 3.9/4

RESEARCH PROJECTS

◆ Research Assistant advised by Prof. Ang Li (06/2024 - present)

Worked on **fine-tuning large language models (LLMs) in multi-agent environments** to analyze and steer internal neural representations for agent honesty and power-seeking behavior. Analyzed the internal neural representations that correspond to high-level social behaviors, including honesty, deception, and power-seeking. Designed template instruction-response prompts to elicit targeted concept activations (e.g., honest vs. dishonest), and applied representation engineering techniques—including **principal component analysis** and **contrastive vector extraction**—to isolate and interpret these internal activations.

Applied **Low-Rank Adaptation (LoRA)** and **control operators** to manipulate concept-specific activations in LLMs, enabling behavior adjustment without full model retraining. These interventions dynamically reshape the model's latent representations during inference, enabling fine-grained control over undesirable behaviors while preserving core task functionality. This framework supports scalable agent honesty verification and demonstrates strong applicability to structured reasoning tasks.

As of **trustworthy AI**, developed EXOgenous Causal reasoning (EXOC), a novel causal inference framework that utilizes auxiliary variables to enhance counterfactual fairness in general ML models. Used counterfactual scenarios to ensure an enhanced accuracy and fairness balance. The paper **"Towards counterfactual fairness through auxiliary variables"** has been accepted to ICLR 2025.

◆ Research Assistant advised by Prof. Yanning Shen, Fairness on Vision Transformers (06/2023 - 10/2023)

Leveraged **adaptive masking** to enhance the fairness-accuracy tradeoff in **Vision Transformers (ViTs)**, a **deep neural network** in **computer vision**. Developed backward-propagation hooked masking in **the attention mechanism** and **contrastive learning** based distance loss, achieved a 6.72% improvement in accuracy compared to leading alternatives with similar fairness. Furthermore, interpretability, efficiency, and computation overhead are carefully crafted and analyzed. The paper "FairViT: Fair Vision Transformer via Adaptive Masking" has been accepted in the European Conference on Computer Vision (ECCV 2024).

◆ Research Assistant advised by Prof. Chuang Gan, Rapper Pose Recognition and Generation (09/2023 - 11/2023)

Collaborated with Prof. Chuang Gan and Mr. Jiaben Chen to refactor and modularize large-scale codebases for OpenPose and TALKSHOW, improving code maintainability and motion detection. Integrated the **YOLOv3 object detection algorithm** to localize human subjects in YouTube videos, and developed a **motion preprocessing pipeline** that extracts, aligns, and normalizes pose sequences. Strengthened expertise in computer vision, multi-person pose estimation, and data-driven action synthesis.

◆ Research Assistant advised by Prof. Qian Wang, Backdoor on Transformers (10/2022 - 06/2023)

Developed a trigger scope limitation strategy to enhance the stealthiness of **backdoor attacks** in **transformers** and manipulated the attention mechanism through "Attention Diffusion" to improve attack flexibility. Implemented these methods, achieving over a 25% improvement in stealthiness and efficiency compared to baseline models. Strengthened expertise in **trustworthy learning** and **foundations of LLMs**. The paper "An Effective and Resilient Backdoor Attack Framework against Deep Neural Networks and Vision Transformers" has been accepted in IEEE Transactions on Dependable and Secure Computing (TDSC 2025).

Extended the proposed **Quality of Experience (QoE)** attack method for both DNN and ViTs, achieving an attack success rate up to 82% higher than baseline methods at low poison ratios and maintaining high QoE in backdoored samples.

◆ Research Assistant co-advised by Dr. Meng Xue and Dr. Xueluan Gong, Dry Eye Disease Detection (01/2023 - 05/2023)

Involved skills in **medical AI application**. Participated in radar-based detection for dry eye disease, offering a convenient and contactless screening method. Analyzed a focal loss-based **Transformer** model in Colab, conducted extensive ablation studies, and reorganized code to implement essential functions like data enhancement, dataset splitting, and model fine-tuning. These efforts culminated in the acceptance of the paper, "SDE: Early Screening for Dry Eye Disease with Wireless Signals" in the Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Ubicomp/IMWUT 2024).

SELECTED PUBLICATIONS

X. Gong*, **B. Tian***, M. Xue, Y. Wu, Y. Chen, Q. Wang. **An Effective and Resilient Backdoor Attack Framework against Deep Neural Networks and Vision Transformers.** (TDSC 2025)

B. Tian, Z. Wang, S. He, W. Ye, G. Sun, Y. Dai, Y. Wu, A. Li. **Towards counterfactual fairness through auxiliary variables.** (ICLR 2025)

B. Tian, R. Du, Y. Shen. **FairViT: Fair Vision Transformer via Adaptive Masking.** (ECCV 2024)

HONORS

Lei Jun Scholarship

Outstanding Student (in 2021, 2022 and 2023)

SKILLS

- Languages: C/C++, Python, SQL, HTML/CSS/JavaScript, MATLAB, Markdown
- Tools: SPSS, VS Code, Jupyter, LATEX, Github, Pytorch, Tensorflow, Conda, Docker
- Research interests: Fairness, Causal Reasoning, AI Security (Backdoor, Data Poisoning), Computer Vision, Vision Transformers, Deep Neural Networks, Natural Language Processing, Large Language Models, AI4Science, Multimodal Models, Federated Learning, Generative AI, Model Inversion, Adversarial Training, Interpretability, Representation Learning