RICHARD RUI CHU

212 W 91st ST, New York, NY | 917-325-4422 | chu.rui@columbia.edu | ruichurich.weebly.com | www.linkedin.com/in/rayrichard22/

EDUCATION

COLUMBIA UNIVERSITY New York, NY
M.S. in Computer Engineering, Department of Computer Science & Electrical Engineering
UNIVERSITY COLLEGE LONDON; FUDAN U - SHANGHAI U SCI TECH London, UK
B.S. in Information Science; Dual in Finance; MSc in Data Science

RESEARCH INTERESTS AND FOCUS

- Trustworthy Artificial Intelligence based on explainable Large Language model inference
- Large-Scale 3D generation in Computer Vision based on Diffusion and NeRF
- Edge-Computing on large model efficiency quantization and adversarial training
- Decision making and reinforcement learning for user willingness
- Multimodal Learning on VR/AR/XR based on sensor fusion and robotics
- Optimizing Federal Learning and distributed Training; Distributed System and Network Optimization

PUBLICATION

- DebiasRAG: Tuning-Free Path to Fair Generation in LLMs through RAG [Under Review NAACL 2025]
- UIBDiffusion: Universal Imperceptible Backdoor Attack for Diffusion Models [Under Review CVPR 2025]
- Less is More: Sparse Watermarking in LLMs with Enhanced Text Quality [Under Reivew ICLR2025]
- Vulnerability of In-Context Learning through adversarial examples [Under Review ICML 2025]
- Controllable Training data generation for 3D models through vLLM generative models [In Submission IJCAI 2025]
- RightCode: Identifying Code Copyright in Large Language Models [In submission ICML 2025]
- Quantization Security: Jailbreak quantization progress while deploying LLMs to Edge [Ongoing on SIGIR 2025]
- Watermark through Retrieval Augmented Generation guided Large Language Model [Submitted to COLM 2024]
- How Chain of Thoughts can be optimized through RAG and In-Context Learning [Ongoing]

PROFESSIONAL EXPERIENCE

TESLA Aug 2023 - Sep 2023
- Cloud Computing Data pipeline establishment in Manufacturing software
- Model Deployment and data streaming pipeline establishment

VIRUFY, BY STANFORD AI HEALTH LAB May 2022 - Sep 2022
- Trained model by tensorflow and deployed on Edge Devices and established AWS web services
- Collected Covid more than 5000 coughing audio dataset and Labelled data for training working & Scheduling with JIRA

BOSCH Apr 2021 - Sep 2021
- Improved system of BOSCH power tools based on data from hardware by 5 percent and tested hardware functions
- Suggested potential improvements on vehicle routing algorithms

RESEARCH AND TEACHING EXPERIENCE

Medical 3D generation based on 2D X-Ray Images Boston, MA
RA Prof. Yingjie Lao, Tufts University
- Generated 3D diagnosing models given X-Ray Image data
- Embedded Text-to-Image Generation so that to have controllable generated data NeRF Training

CLAUDIUS LEGAL INTELLIGENCE AI, NLP lab at Princeton University Boston, MA
Research Assistant with Prof. Melanie Weber, Harvard
- Cloud Retrained Legal paper and materials based on previous models

CREATIVE MACHINES LABORATORY, CUDA/ OPENGL New York, NY
Researcher with Prof. Hod Lipson & Dr. Phillipe, Columbia
- Parallel-computed Simulation Environment by IssacGym to accelerate robots training performance through RL
- Accelerated Laurenz-curve calculation on Cloud GPU by designing CUDA Kernel

Reinforcement Learning FOR AUTO DRIVE ROUTING IN SIMULATION ENVIRONMENT New York, NY
Research Assistant with Prof. Sharon Di, Columbia
- Built simulation environment with Unity to train car to avoid obstacles and applied Transformer model
- Implemented DQN and DDPG to train model by Pytorch and CUDA, and improved model by fine tuning

COMPUTER VISION IN NETWORK-FLUCTUATION-BASED WIND PREDICTION New York, NY
RA Prof. Gil Zussman& Igor Kadota, Columbia & Northwestern
- Trained ImageNet Computer Vision to do classification and predict wind map generated by mmWave
- Reinforcement Learning for MIMO beamforming base stations

VIRTUAL REALITY DESIGN FOR BRITISH MUSEUM London, United Kingdom
Project Leader with Prof. Josep Grau-Bove, UCL
- Crowd-Sourcing Image data from tourists to generate point cloud for buildings and collections
- Designed improved 3D virtual Reality tour using Unity and 3D sensor fusion

Visual Interfaces to Computers New York, NY
Teaching Assistant of Prof. John Kender, Columbia University

RICHARD RUI CHU

212 W 91st ST, New York, NY | 917-325-4422 | chu.rui@columbia.edu | ruichurich.weebly.com | www.linkedin.com/in/rayrichard22/

EDUCATION

COLUMBIA UNIVERSITY New York, NY
M.S. in Computer Engineering, Department of Computer Science & Electrical Engineering
UNIVERSITY COLLEGE LONDON; FUDAN U - SHANGHAI U SCI TECH London, UK
B.S. in Information Science; Dual in Finance; MSc in Data Science

RESEARCH INTERESTS AND FOCUS

- Trustworthy Artificial Intelligence based on explainable Large Language model inference
- Large-Scale 3D generation in Computer Vision based on Diffusion and NeRF
- Edge-Computing on large model efficiency quantization and adversarial training
- Decision making and reinforcement learning for user willingness
- Multimodal Learning on VR/AR/XR based on sensor fusion and robotics
- Optimizing Federal Learning and distributed Training; Distributed System and Network Optimization

PROFESSIONAL EXPERIENCE

TESLA
Data Streaming Engineer Aug 2023 - Sep 2023
- Cloud Computing Data pipeline establishment in Manufacturing software
- Model Deployment and data streaming pipeline establishment

VIRUFY, BY STANFORD AI HEALTH LAB
Software Engineer/ DevOps May 2022 - Sep 2022
- Trained model by tensorflow and deployed on Edge Devices and established AWS web services
- Collected Covid more than 5000 coughing audio dataset and Labelled data for training working & Scheduling with JIRA

BOSCH
Embedded Engineer Apr 2021 - Sep 2021
- Improved system of BOSCH power tools based on data from hardware by 5 percent and tested hardware functions
- Suggested potential improvements on vehicle routing algorithms

PUBLICATION

- DebiasRAG: Tuning-Free Path to Fair Generation in LLMs through RAG [Under Review NAACL 2025]
- UIBDiffusion: Universal Imperceptible Backdoor Attack for Diffusion Models [Under Review CVPR 2025]
- Less is More: Sparse Watermarking in LLMs with Enhanced Text Quality [Under Reivew ICLR2025]
- Vulnerability of In-Context Learning through adversarial examples [Under Review ICML 2025]
- Controllable Training data generation for 3D models through vLLM generative models [In Submission IJCAI 2025]
- RightCode: Identifying Code Copyright in Large Language Models [In submission ICML 2025]
- Quantization Security: Jailbreak quantization progress while deploying LLMs to Edge [Ongoing on SIGIR 2025]
- Watermark through Retrieval Augmented Generation guided Large Language Model [Submitted to COLM 2024]
- How Chain of Thoughts can be optimized through RAG and In-Context Learning [Ongoing]

RESEARCH EXPERIENCE

Medical 3D generation based on 2D X-Ray Images Boston, MA
RA Prof. Yingjie Lao, Tufts University
- Generated 3D diagnosing models given X-Ray Image data
- Embedded Text-to-Image Generation so that to have controllable generated data NeRF Training

CLAUDIUS LEGAL INTELLIGENCE AI, NLP lab at Princeton University Boston, MA
Research Assistant with Prof. Melanie Weber, Harvard
- Cloud Retrained Legal paper and materials based on previous models

CREATIVE MACHINES LABORATORY, CUDA/ OPENGL New York, NY
Researcher with Prof. Hod Lipson & Dr. Phillipe, Columbia
- Parallel-computed Simulation Environment by IssacGym to accelerate robots training performance through RL
- Accelerated Laurenz-curve calculation on Cloud GPU by designing CUDA Kernel

Reinforcement Learning FOR AUTO DRIVE ROUTING IN SIMULATION ENVIRONMENT New York, NY
Research Assistant with Prof. Sharon Di, Columbia
- Built simulation environment with Unity to train car to avoid obstacles and applied Transformer model
- Implemented DQN and DDPG to train model by Pytorch and CUDA, and improved model by fine tuning

COMPUTER VISION IN NETWORK-FLUCTUATION-BASED WIND PREDICTION New York, NY
RA Prof. Gil Zussman& Igor Kadota, Columbia & Northwestern
- Trained ImageNet Computer Vision to do classification and predict wind map generated by mmWave
- Reinforcement Learning for MIMO beamforming base stations

Visual Interfaces to Computers New York, NY
TA Prof. John Kender, Columbia University
- Guided the student groups on varieties of computer vision projects

Ph.D. Personal Statement for Columbia University

My life goal is: *Live to push human beings forward*. With my industry experience at Tesla which focuses on vision AI, I have realized that there are still a lot of concerns need to be solved in current AI development, and I live to accelerate the generating world revolution.
Thus, to realize my life goal, my research interest would be: **Trustworthy, controllable, explainable, and resource-friendly generation for Large Models such as language models as well as vision models** like diffusion and 3D models like NeRF, so that to establish a controllable large-generation world based on explainable statistical learning.

My passion in this area started from my early research for constructing Virtual Reality for the British Museum at University College London, where I collected data from point clouds and illustrated the view through Unity with the team of Google. Since then, I have imagined if the point cloud could be generated from clean noise. Thus, during my Master's study, when I first played with Transformers, such as ViT and GPT2, I always dreamed about a generative world. Before the bombing of GPT3, I conducted research on Reinforcement Learning for vision Auto-driving in smart cities, as well as 3D simulation environment establishment through IssacGym for parallel computing training acceleration, and implementing ResNet to identify weather wind influence to network systems of edge devices. All these prior experiences prepared me well in a perfect research sense and solid technical background.

With the exposure to emerging large generative models, **I moved everything I had equipped into the generative domain.** Since the Transformer has dominated the industry, I mostly focused on inferencing and reasoning method exploration of the models to improve human-AI interactions in a trustworthy, explainable, and resource-friendly way. One of my first steps was *DebiasRAG*, where I used Retrieval Augmented Generation to enhance the fairness of LLM outputs by generating Biased output and reversing them as an RAG injection to make a fair generation. Based on the nice experimental output, I have questioned myself why RAG, as a prompt searcher, could have an additional fairness improvement impact even on unseen data. Thus, I dived into In-Context Learning to understand how theoretically the model handles the prompts. During the time better understanding the experimental RAG results in an explainable way, I found that the robustness of In-Context Learning should be awarded, and thus am working on a theoretical proof of vulnerabilities among ICL, with the methodology of simulating the inferencing tasks into nonlinear function and apply Projected Gradient Descent onto it. Meanwhile, one of my other works experimented with how In-Context Learning can influence the Chain of Thoughts, and am now implementing RLHF onto ICL progress inspired by CoT. By researching LLM inference

theory, on the other direction, I conducted Large Model Quantization through skipping layers and am currently experimenting with it to make the AI inference more resource-friendly, which is increasingly important in bigger LLMs like GPT5 and will be helpful on edge deployment and federal learning. In addition, I have done copyright code feature extraction to improve LLM copyright identification and watermarking LLMs. One of my goals in LLM research currently would be to make the LLMs able to be taught, trained, and instructed as easily as humans in a trustworthy way so that the human education industry can be merged with AI tuning directly.

Generative Models are developing fast, not only in the language domain but also in vision areas, such as 2D Diffusion and 3D NeRF. Thus, I have extended my methods and computer vision background into vision models. One of my works *UIBDiffusion* has implemented a universal adversarial perturbation as an imperceptible backdoor, revealing the weakness of the current diffusion progress. Furthermore, I have experimented with 3D generation through NeRF based on 2D X-ray images and verified if the 3D NeRF outcomes can still be correctly classified and improved the loss based on the misclassifying rate. All my latest research can be concluded as a chain: how we can establish a controllable Text-Image-3D-Video-VR multimodal generation, which will be exploding in new areas like 3D-printing-distributed manufacturing, smart cities, even filming and financial areas, etc. Thus, I believe all these works get me well-prepared for my further PhD level research.

However, it might be surprising that my actual motivation to conduct a Ph.D. level study is not only for achieving detailed productive outcomes, which is the basic requirement for a Ph.D. student, but also to keep looking up to the sky to reveal the counterintuitive truth. Starting from a kid, I always questioned myself why I needed to study, and was finally convinced that I study to figure out how the world actually works. Thus, I was trained on a wide range of datasets by undergraduate study, including Information Science, Economics, etc; and was fine-tuned on Machine Learning and Computer Engineering tasks by Columbia during my Master's study so that I could be equipped with meta-learning on a wide range of domains. It is important because one of the basic research paths is to be inspired by other domains and then implement with own SOTA: for example inspired by biology, humans have designed aircraft wings and neural networks; and implementing attention on RNN forms Transformer. I have already tried this level of research by submitting the papers like *UIBDiffusion*. Beyond these experiment-based researches, I am conducting explainable research like revealing in-context learning vulnerability so that I can dive deeper into theoretical exploration to reveal unknowns — As a second step I hope I can explain everything properly with proper theory.
But it is still not enough because I dream big. From a high-level perspective, People are always researching some general areas: human society such as Economics to learn

ourselves; productive tools such as Computer science to lead a technical revolution with explosive productivity, and Natural sciences such as Physics, as well as philosophy so that to reveal the truth of the universe. Recalling my life goal is to push human beings forward, I hope my research will further clarify what is the actual rule of the universe, just like Stephen Hawking suggested no-boundary proposal; and then propose a theorem or a technique that can generate the next era: for example, I deeply believe that we are living in a virtual world, agreed by Elon Musk — and my research in generative world will finally make the human beings become God in a new virtual universe so that prompt the current human into a next Kardashev Scale civilization which could control the time and all nature resources. These plans require a general understanding of a variety of areas, which I not only got well prepared by my previous experience both from studying and working with startups from Stanford and Princeton labs, but also will be solidified by cooperating with researchers in other departments at Columbia. Ideally, I hope that my research impact will not only last a 5-year long Ph.D. journey, but also become my life-long research task in start-up companies with colleagues I worked with during my PhD — my previous experience let me understand the authentic works, but I will keep proposing hypotheses to question them and carefully testing the result to find out counterintuitive truths, Just like Feynman reconstructed quantum electrodynamics. Witnessing John J. Hopfield achieve the latest Nobel Prize in Physics, the path of being impactful through conducting AI research has never been as clear as today.
Thus, my general plan for my Ph.D. level research will be to spend the first two years on producing research and industrial impactful SOTA works, and the rest of the time will be contributing to theoretical breakthroughs, starting from my domain and then expanding to varieties of areas, for example, biomedical domain as what I have done in NeRF 3D generation for X-Rays, and social science area that could establish a quantified connection between distributed computing power and blockchain to properly allocate federal learning in a financial way; and also smart city domains, etc. At the end of my PhD journey, I believe I am running on the path of generating a new universe till the end of my life.

In a word, my PhD will be the start of my research journey, and I hope I can reveal the explainable truth by simulating them into a generative world, and then lead mankind into a new generative era years later. As a graduated Columbia Lion, I am very glad to find that some amazing advisors align with my research interests greatly. I believe I will be impactful starting from Columbia, which has a rich history of realizing crazy ideas. Thank you for your consideration.

As a graduated Columbia Engineering student, I am very glad that I found advisors that fit my research interest really well this year and thus I am deeply eager to do research back to Columbia.

As a PhD applicant, before Claiming my research interest, I would like to propose my life goal: Live to push human beings forward a little bit. To realize it, one of my plans is to illustrate a generative world for new human beings, not only from the technical side, but also from the society perspective: how the new world could be established in a better way.

Thus, my research interest is: Large-Scale controllable Trustworthy AI generation based on Human-AI interaction. To be more detailed, ranging from statistical learning as a first, to Large Scale VR-3D generation based on diffusion models and NeRF, Human-AI interaction optimization for Language Models and Vision Models, and finally an easy to control and user-friendly large-scale generation world for the entire mankind.

I do get well prepared for my research goal. Since my early study at University College London, I optimized the Virtual Reality system of the British Museum online with the team of Google. The VR system played a key role during the pandemic since visitors can visit online and is still being used till now. After I moved to Columbia, I conducted a 3D simulation environment with IssacGym for Robot training, where the robots can be substituted as anything in a 3D generation world. Based on this experience, I worked with the civil lab to conduct Reinforcement Learning on Auto-Driving training. Those experiences gave me a solid foundation in computer vision training, and thus I worked with Prof. Kender on vision systems.

A qualified PhD student should have two sense: use tool and think about how actually the world is running
My plan for my PhD study, there are several steps:

I believe virtual world will help real worlds and finally all mankind will live in dreaming good virtual world