Personal Statement

Brian Nlong Zhao
briannlongzhao@gmail.com

My research in artificial intelligence - in particular computer vision and graphics - has stemmed from a curiosity in how machines can interpret and reproduce our world visually. **My interests center around helping vision models understand complex distributions and dynamics of our world, especially through data-centric approaches.** Current vision models, while effective in fundamental tasks like static image classification and segmentation, often fail when it comes to capturing complex distributions and understanding the physical properties and behaviors of objects over time. This gap becomes evident when models reconstruct or generate physically inconsistent motions or unrealistic object interactions in dynamic contexts such as video and animated 3D renderings. I believe the insufficiency of high-quality data is one of the most critical factors that contribute to this limitation, especially because of the challenges manually curating large-scale, high-quality annotations brings, particularly for dynamic tasks that require detailed annotations of motion, interactions, and physical properties over time. Moreover, data-efficient physics-integrated representations offer another key solution. Current models often rely on static or overly simplified representations, which fail to incorporate essential physical properties such as deformation, elasticity, and motion under external forces. Therefore, my research agenda is to bridge this gap by designing data-centric approaches that allow for more accurate and reliable performance in tasks that require an understanding of the dynamics of objects and scenes beyond a surface level, such as inverse rendering of dynamic and articulated objects and motion generation of humans, animals, and articulated objects.

My research journey has been a gradual exploration, beginning with foundational computer vision tasks such as image super-resolution and feature matching, and evolving toward more complex challenges like inverse rendering and motion reconstruction. I started during the pandemic when I was taking classes remotely from my home in Shanghai. I sought opportunities to work on AI research and connected with Professor Lin Zhang and Professor Yu Wang at Tongji University, where I contributed to projects in low-level computer vision problems, including camera calibration [1], image feature matching [5], and super-resolution [8]. I primarily focused on collecting, processing, and organizing various datasets such as car surround-view and remote sensing data, as well as enhancing the accuracy and efficiency of algorithms. This experience provided me with a strong technical foundation in computer vision and machine learning, particularly in collecting, organizing, and analyzing data. It also enhanced my understanding of how machines learn from data and fueled my passion for addressing real-world challenges in building efficient and scalable datasets for vision systems.

Upon returning to campus in my senior year, I joined the USC iLab supervised by Professor Laurent Itti, and worked closely with Dr. Yunhao Ge on generative AI for vision applications. I was drawn to data-centric methods and participated in projects on synthetic dataset generation for object detection and segmentation [2, 3]. We used a cut-and-paste technique combined with text-to-image generation models to produce synthetic data at an infinite scale, demonstrating significant improvements in detection and segmentation metrics beyond state-of-the-art models. As a core contributor, I took the lead in designing and implementing the prompt generation workflow and a novel EM-based iterative segment filtering algorithm for cut-and-paste. This work garnered over 50 citations this year, which encouraged me and strengthened my commitment to advancing computer vision research and its applications.

Driven to dive deeper into research, I extended my studies for one more year for a Master's degree at USC, and I secured an internship at Microsoft Research Asia, where I explored AI's potential in the realm of healthcare. Guided by Dr. Dongsheng Li and Dr. Xinyang Jiang, I led a project to create a million-scale dataset for radiology report generation [6]. Using GPT-based methods inspired by projects like LLaVA [4], we constructed a benchmark dataset that reflects real-world clinical scenarios, where radiologists construct reports while considering context like structured templates or patient medical history. We also developed an efficient cross-attention mechanism that adapts domain-specific vision encoders with pretrained multimodal LLMs, as well as a novel pathological guidance module, optimizing both medical context understanding and clinical efficacy. Leading this project was a unique interdisciplinary experience for me, as it involved frequent communication with medical professionals from local hospitals, regular visits for meetings, and observing real clinical diagnostic scenarios, which grounded my work in real-life experiences of people. This large-scale project has resulted in a paper submission to ICLR 2025, marking my first time submitting to a top-tier AI conference as the first author.

During my Master's, my growing interest in generative AI led me to initiate a project at USC iLab. I was particularly interested in how generative models can handle diverse data of complex distribution and create flexible outputs while maintaining fidelity. With personalized generation tasks as the entry point, I aimed to demonstrate the importance of modeling the distribution of data, rather than instance-specific personalization. Under the guidance of Professor Itti, Dr. Ge, and Dr. Vibhav Vineet from Microsoft Research, I developed DreamDistribution [7], a method that advances personalized image and 3D generation by capturing a broad range of visual attributes from reference images. Inspired by how dreams reassemble elements of reality into new perspectives, DreamDistribution uses prompt distribution learning to model complex features like style, pose, and texture, enabling in-distribution instance generation with appropriate variations. This is especially useful in tasks like text-to-image or text-to-3D generation, where balancing diversity and fidelity is key. At the same time, I was very interested in the potential of applying prompt distribution learning to other domains such as text-based visual classification and prediction. To explore the possibilities of the applications, I collaborate with Professor Wang at Tongji University to adapt similar ideas to tackle the animal pose estimation problem. The two projects, which resulted in a submission to ICLR 2025 as first-author and CVPR 2025 as co-author respectively, led me think more deeply about how to model real-world distributions related to concepts such as physics and motion via data-centric approaches.

Inspired by Professor Jiajun Wu's works on modeling physics and motion through computer vision models, I joined his Vision and Learning Lab at Stanford University, working closely with Dr. Shangzhe Wu. My current research focuses on modeling real-world motion distributions of articulated objects, such as animals, through inverse rendering - transforming 2D images and videos into 3D and dynamic 4D reconstructions - and generating realistic motion based on them. I developed a large-scale video scraping and processing pipeline that automatically gathers videos of specific categories and applies a comprehensive workflow, including segmentation, tracking, as well as depth, flow, occlusion estimation, feature extraction, object-centric filtering, etc. This process generates an expansive dataset that enables inverse-rendering models to derive 3D shapes and motions from 2D media at scale. Currently, the dataset includes millions of high-quality, object-centric frames capturing animal motions, which I believe no other dataset has achieved at this scale, and it continues to expand. Fortunately, this work has led to a first-author publication at NeurIPS 2025 Datasets and Benchmarks track, and I am now excited to continue with follow-up research that leverages our dataset for advancing animal motion generation.

Now as a first-year PhD student at UIUC working with Professor James Rehg, I am exploring how generative models can address data scarcity in scientific domains. Our current focus is on eye-tracking data, where most existing datasets are restricted to hospitals and clinical institutions. We propose using diffusion models to learn the distribution of limited public datasets and to generate realistic synthetic eye-tracking data that reflects human attention patterns. In the long term, this line of research could enable deep learning methods to advance a wide range of downstream applications: from supporting the diagnosis of autism and other neurological disorders, to analyzing cognitive load and attention, to improving user interface design, user experience, and even advertising effectiveness. Ultimately, I am motivated by the possibility that generative models can bridge data gaps and contribute to impactful applications in healthcare, defense, and everyday human-centered technologies.

References

[1] Yang Chen, Lin Zhang, Ying Shen, **Brian Nlong Zhao**, and Yicong Zhou. "Extrinsic Self-Calibration of the Surround-View System: A Weakly Supervised Approach". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 1622–1635. DOI: 10.1109/TMM.2022.3144889.

[2] Yunhao Ge, Jiashu Xu, **Brian Nlong Zhao**, Laurent Itti, and Vibhav Vineet. "EM-Paste: EM-guided Cut-Paste with DALL-E Augmentation for Image-level Weakly Supervised Instance Segmentation". In: *arXiv preprint arXiv:2212.07629* (2022).

[3] Yunhao Ge, Jiashu Xu, **Brian Nlong Zhao**, Neel Joshi, Laurent Itti, and Vibhav Vineet. "Beyond Generation: Harnessing Text to Image Models for Object Detection and Segmentation". In: *arXiv preprint arXiv:2309.05956* (2023).

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual Instruction Tuning". In: *NeurIPS*. 2023.

[5] Yizhang Liu, **Brian Nlong Zhao**, Shengjie Zhao, and Lin Zhang. "Progressive Motion Coherence for Remote Sensing Image Matching". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–13. DOI: 10.1109/TGRS.2022.3205059.

[6] **Brian Nlong Zhao**, Xinyang Jiang, Xufang Luo, Yifan Yang, Bo Li, Zilong Wang, Javier Alvarez-Valle, Matthew P. Lungren, Dongsheng Li, and Lili Qiu. "Large Multimodal Model for Real-World Radiology Report Generation". Dec. 2023. URL: https://www.microsoft.com/en-us/research/publication/large-multimodal-model-for-real-world-radiology-report-generation/.

[7] **Brian Nlong Zhao**, Yuhang Xiao, Jiashu Xu, Xinyang Jiang, Yifan Yang, Dongsheng Li, Laurent Itti, Vibhav Vineet, and Yunhao Ge. "Dreamdistribution: Prompt distribution learning for text-to-image diffusion models". In: *arXiv preprint arXiv:2312.14216* (2023).

[8] Cairong Zhao, Shuyang Feng, **Brian Nlong Zhao**, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. "Scene Text Image Super-Resolution via Parallelly Contextual Attention Network". In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, 2021, pp. 2908–2917. ISBN: 9781450386517. DOI: 10.1145/3474085.3475469. URL: https://doi.org/10.1145/3474085.3475469.