Guanglei (Ian) Zhu

guanglez@andrew.cmu.edu | 412-598-4541 | LinkedIn | Website

Education

Carnegie Mellon University - School of Computer Science Pittsburgh, PA
Master of Science in Computer Vision, GPA: 3.95/4.00 Dec. 2024
- **Relevant Courses:** Multimodal Foundation Models, Large Language Models, Visual Learning and Recognition

University of Toronto Toronto, Canada
Honours Bachelor of Science in Computer Science, GPA: 3.87/4.00 May 2023
- **Relevant Courses:** Computer Vision, Deep Learning, Algorithms & Data Structures, Operating Systems, Computer Networks, Software Design, Database Systems, Computer Organization
- **Awards:** In-Course Scholarship (2020), Dean's List Scholar (2019 - 2021)

Work Experience

TikTok San Jose, CA
Machine Learning Engineer Intern May 2024 - Aug. 2024
- Pretrained the Content Violation Prediction Model (CVPM) by integrating a Perceiver Sampler module to improve temporal understanding, boosting precision by 10% and recall by 8% for the updated foundation model
- Collaborated with cross-functional teams to integrate content and user-level features features, training the first CVPM Ensemble model and achieving a 8% increase in precision and recall over the latest CVPM model
- Responded to a high-exposure photo slice violation trend by deploying a detection model trained on synthesized data

Vector Institute Toronto, Canada
Research Intern, Advised by Prof. Animesh Garg Apr. 2022 – May 2023
- Proposed a differentiable rendering pipeline for hand-object pose estimation by optimizing 3D renders against 2D images, achieving a 30% reduction in mesh error compared to state-of-the-art methods
- Developed an end-to-end pipeline to predict hand-object pose trajectories from RGB videos by integrating Faster R-CNN for object detection, PointRend for segmentation, and refinement using pre-trained models
- Published HandyPriors: Physically Consistent Perception of Hand-Object with Differentiable Priors at **ICRA 2024**

Projects

Run-Length Tokenization for Faster Video Transformers Jan. 2024 - May 2024
CUBE Lab by Prof. Laszlo Jeni, Carnegie Mellon University Pittsburgh, PA
- Developed Run-Length Tokenization (RLT), a content-aware, dataset-agnostic method to accelerate video transformers by efficiently processing repeated input patches, reducing redundant data and minimizing overhead
- Implemented RLT in Epic Kitchen 100 Multi-Instance Retrieval task, reducing fine-tuning time by 30% compared to baseline without performance drop
- Published Don't Look Twice: Faster Video Transformers with Run-Length Tokenization at **NeurIPS 2024 Spotlight**

Self-Bias Amplifies in Large Language Model Oct. 2023 - April 2024
Li's Lab by Prof. Lei Li, Carnegie Mellon University Pittsburgh, PA
- Analyzed self-feedback in LLMs, identified prevalent self-bias (tendency to favor its own generation) across translation, constrained text generation, and mathematical reasoning in all popular LLMs (e.g. GPT4, LLaMA2, Mistral MoE)
- Explored self-bias in the LLM self-rewarding pipeline, discovered that larger sample sizes increased bias and distance skewness, exacerbating self-bias during training
- Published Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement at **ACL 2024 Oral**

Skills

- Programming Language: Python, C/C++, Java, PostgreSQL, JavaScript
- Libraries & Tools: PyTorch, PySpark, Hadoop, OpenCV, Scikit-learn, Pandas, Numpy, Git, Docker, Slurm, Linux
- Models: Transformer, BERT, CLIP, LLAVA, LLaMA, LSTM, CNN, Decision Tree, Logistic Regression, Neural Network