Jialuo Li

☎ 470-265-8166 | ✉ jli3671@gatech.edu | 🌐 jialuo-li.github.io | Jialuo-Li

Education

College of Computing | Georgia Institute of Technology Aug. 2025 - Present
Graduate student in Computer Science

Yao Class | Tsinghua University Sept. 2021 - July. 2025
Bachelor of Engineering in Computer Science and Technology. GPA 3.84/4.0

Employment

Georgia Institute of Technology Aug. 2025 - Present
Research Assistant | Advisor: Humphrey Shi

Microsoft Research Asia Sep. 2024 - May. 2025
Research Intern | Mentor: Bin Li

New York University Feb. 2024 - July. 2024
Research Intern | Advisor: Saining Xie

Research Interest

Humans exhibit an extraordinary capacity to integrate information from multiple sensory modalities, such as vision, auditory, and tactile inputs, to navigate and interpret their environments with remarkable efficiency. This multimodal integration leverages the complementary strengths of each sensory channel, facilitating a coherent and comprehensive understanding of complex surroundings. Inspired by this cognitive prowess, my long-term objective is to develop AI systems that emulate human-like multimodal synthesis, thereby enhancing robustness and adaptability in both generative and understanding tasks. Specifically, I focus on **multimodal generative models**, ranging from text-to-image synthesis to Multimodal Large Language Models (MLLMs) and unified models.

Publications and Manuscripts

(* stands for equal contribution.)

[1] *(Under Review)* Jialuo Li*, Fengzhe Zhou*, Jiannan Huang*, Deva Ramanan, Humphrey Shi. "PAI-Bench: A Comprehensive Benchmark For Physical AI." PDF
TL;DR | PAI-Bench is a new benchmark designed to evaluate Physical AI capabilities across video generation and understanding. The study finds that while current models produce high-quality visuals, they lack the physical common sense and reasoning required to truly understand real-world dynamics.

[2] *(Under Review)* Jitesh Jain, Jialuo Li, Zixian Ma, Jieyu Zhang, Chris Dongjoo Kim, Sangho Lee, Rohun Tripathi, Tanmay Gupta, Christopher Clark, Humphrey Shi. "SAGE: Training Smart Any-Horizon Agents for Long Video Reasoning with Reinforcement Learning." PDF
TL;DR | SAGE introduces a human-inspired agentic framework that replaces resource-heavy frame processing with iterative reasoning and a diverse toolkit for efficient long video understanding. By leveraging a novel synthetic data pipeline and a specialized reinforcement learning strategy, SAGE significantly outperforms state-of-the-art models on open-ended tasks within the new SAGE-Bench.

[3] *(Under Review)* Min Shi, Xiaohui Zeng, Jiannan Huang, Yin Cui, Francesco Ferroni, Jialuo Li, Zhaoshuo Li, Yogesh Balaji, Haoxiang Wang, Tsung-Yi Lin, Xiao Fu, Yue Zhao, Chieh-Yun Chen, Ming-Yu Liu, Humphrey Shi. "DuetGen: Towards General Purpose Interleaved Multimodal Generation." PDF
TL;DR | DuetGen enables general-purpose interleaved multimodal generation by fusing a pretrained MLLM with a video-pretrained DiT, avoiding the cost of unified pretraining. Leveraging a curated 298k-sample dataset, it significantly outperforms open-source baselines in visual fidelity and consistency across multiple benchmarks.

[4] *(Under Review)* Jialuo Li, Bin Li, Jiahao Li, Yan Lu. "DIvide, then Ground: Adapting Frame Selection to Query Types for Long-Form Video Understanding." PDF
TL;DR | DIG optimizes long-form video understanding by distinguishing between "global" and "localized" queries, applying uniform sampling for the former and a specialized retrieval pipeline for the latter. This training-free approach consistently outperforms existing baselines across multiple benchmarks, robustly enhancing LMM performance even with high frame counts

[5] *(CVPR 2025)* Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, Saining Xie. "Science-T2I: Addressing Scientific Illusions in Image Synthesis." PDF
TL;DR | Science-T2I addresses the tendency of generative models to produce ”scientific illusions” by introducing a dataset of over 20k expert-annotated image pairs rooted in physics, chemistry, and biology. Leveraging the new SciScore reward model and a two-stage fine-tuning framework, the approach significantly improves the scientific realism of generated images compared to state-of-the-art baselines.

[6] *(COLM 2024)* Jingzhe Shi, Jialuo Li, Qinwei Ma, Zaiwen Yang, Huan Ma, Lei Li. "CHOPS: Chat with Customer Profile Systems for Customer Service with LLMs." PDF
TL;DR | CHOPS introduces a Classifier-Executor-Verifier framework that enables LLMs to accurately query databases and execute system operations for customer service. Validated on the new CPHOS-dataset, this approach achieves 98% accuracy while significantly reducing costs by strategically combining small and large models.

Services

Teaching Assistant in Georgia Tech	Aug. 2025 - Dec. 2025
Class monitor of Yao Class	Sept. 2022 - Sept. 2023
President of the IIIS Student Union Organization Group	Sept. 2022 - Sept. 2024
CPHO-S co-founder, former tech group leader, council member.	Sept. 2021 - Sept. 2022

Awards and Scholarships

Scholarships

Social Worker Merit Scholarship of Tsinghua	Oct. 2022
Social Worker Merit Scholarship of Tsinghua	Oct. 2023
Second-Class Freshmen Scholarship of Tsinghua	Nov. 2021

Competitive Physics in High School

1st Prize in National High School Physics Olympics Competition (**10th Place Nationwide**)	Nov. 2020
National Team member of China for Asian Physics Olympiad (APhO)	Feb. 2021 - May. 2021