Ph.D. Dissertation Defense

Candidate:	Yangyang Yu
Degree:	Doctor of Philosophy
School/Department.:	Interdisciplinary / Data Science
Date:	Monday, May 5, 2025
Time:	11:00 – 12:30 pm
Location:	https://stevens.zoom.us/j/4260484606
Title:	Aligning Multi-modal Object Representations to Human Cognition
Chairperson:	Dr. Jordan Suchow, Information Systems, School of Business
Committee Members:	Dr. Zining Zhu, Computer Science, School of Engineering & Science Dr. Denghui Zhang, Information Systems, School of Business Dr. Rong Liu, Information Systems, School of Business Dr. Mark Ho, Department of Psychology, New York University Dr. Joshua Peterson, Computing and Data Science, Boston University

Abstract

Cognitive abilities develop as individuals accumulate life experiences from interacting with their complex environments. This progression is influenced by various events and objects encountered along the way. However, understanding how humans formulate perceptions in response to diverse external stimuli poses ongoing challenges in behavioral research, especially within the constraints of traditional laboratory settings. These environments often suffer from limited budgets, reliance on simplistic features, and the inability to fully capture the interrelationships among stimuli. Recent advancements in deep learning and artificial intelligence, along with the expansion of web data and online crowdsourcing platforms, are creating new possibilities for more accurately aligning human cognition with a broad spectrum of objects. These technologies offer significant improvements in research quality and efficiency by enabling better prediction and understanding of human perceptual responses to various objects. Two primary approaches have emerged to advance this field. First, applying machine learning and deep learning techniques—such as high-dimensional feature extraction, tensor fusion methods, attention mechanisms, and active sampling strategies—enables richer, more nuanced representations of objects and their interactions. Second, a new generation of generative agents leveraging Large Language Models (LLMs) and Vision-Language Models (VLMs) significantly enhances agents' comprehension of multimodal environmental information, facilitates the development of sophisticated memory systems that closely mirror human cognition, and empowers agents to perform complex actions. Particularly within domains requiring deep professional expertise, such as the fashion industry, these approaches offer substantial improvements in accurately presenting, analyzing, and interpreting human behavior to support informed decision-making.

In this thesis, we demonstrate three advanced methodologies that markedly enhance both the quality and efficiency of aligning multimodal perception with human cognitive processes. The first approach introduces an innovative multi-modal fusion framework, leveraging matrix factorization enriched with deep learning-generated features, to effectively align human perceptual outcomes with machine-generated multi-modal stimuli. Specifically, this framework facilitates accurate prediction of human behaviors, exemplified through applications like predicting first impressions from facial images combined with psychological attributes. This second explores the use of active learning strategies to efficiently identify the most informative samples, constructing a compact yet highly predictive training dataset for multi-modal data fusion. Active learning addresses data sparsity challenges by accurately estimating full response distributions from limited selective samples. The last research introduces **FashionAgent,** an innovative AI generative agent framework specifically designed for fashion industry tasks, featuring automated extraction and explanation of outfit features using Vision-Language Models. It also uniquely integrates human-inspired template memory theory to significantly enhance judgment accuracy, output quality, and processing efficiency.