Research Statement

LI Jiannan
School of Computing and Information Systems, Singapore Management University
Tel: (65) 68264834; Email: jiannanli@smu.edu.sg
18 (Day) 12 (Month) 2025 (Year)

Background

The interaction paradigm with intelligent agents today (e.g. robots, chatbots) primarily uses text and graphical interfaces. However, effective human communication often involves a richer set of modalities (e.g. vision, language, audio) and combine them fluidly. For example, people pay attention to the subtle body languages of others as signals of approval and learn better with textbooks that visualize certain key concepts through illustrations and diagrams. Theories of embodied cognition and multimedia learning note that appropriate combinations of multiple modalities are conducive to communication and learning. My research thus aims to expand the bandwidth of human-agent communication through effectively integrating multiple modalities in both input and output.

While effective multi-modal interaction has been a long-standing research problem, prior solutions were often found brittle due to their reliance on hand-coded rules or small datasets. My work takes the unique approach of guiding the strong generalization abilities of AI foundation models with human behavior insights. More concretely, I identify theories from communication studies, social psychology, and learning science, for constructing input and output modality combinations that reveal hidden human intents and clearly convey the intended messages (e.g. knowledge to learn, instructions to follow) to human users. As a next step, these theories are generalized to a wide range of possible scenarios using Large Language Models and integrated with multimodal perception and generation capabilities. My work is more specifically focused on the following two spaces: (1) developing collaborative robots that learn human intents through multimodal signals, and (2) building intelligent training and learning systems with multimodal understanding and feedback.

Research Areas

Collaborative Robots that Learn Human Intents through Multimodal Signals

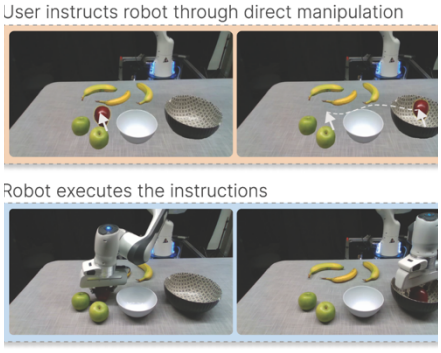

Figure 1 *We introduce the direct manipulation of images as a paradigm for providing instructions to a robot. Depicted in the bottom left are a series of instructions that a user is giving the robot by manipulating the fruits. The top shows one trajectory of direct manipulation.*

Theme 1: Robot Programming through Direct Image Manipulation and Language

Foundation models are rapidly improving the capability of robots in performing everyday tasks autonomously such as meal preparation, yet robots will still need to be instructed by humans due to model performance, the difficulty of capturing user preferences, and the need for user agency. Robots can be instructed using various methods---natural language conveys immediate instructions but can be abstract or ambiguous, whereas end-user programming supports longer-horizon tasks but interfaces face difficulties in capturing user intent.

In this work, we propose using direct manipulation of images as an alternative paradigm to instruct robots, and introduce a specific instantiation called ImageInThat which allows users to perform direct manipulation on images in a timeline-style interface to generate robot instructions (Figure 1). Through a user study, we demonstrate the efficacy of ImageInThat to instruct robots in kitchen manipulation tasks, comparing it to a text-based natural language instruction method. The results show that participants were faster with ImageInThat and preferred to use it over the text-based method.

Theme 2: Robots that Understand Human-to-Human Interactions to Provide Help
Remote assistance through robotic telepresence could involve both navigation and information challenges, particularly in one expert to multiple workers situation. In this study, we proposed a novelty language-driven interface to facilitate remote collaboration through telepresence robots. Through operations and maintenance expert interviews and a scenario simulation study, we identified key pain points in executing one-expert-multiple-workers remote guidance using the telepresence robot and proposed five design goals with corresponding features. These features were integrated into a standard telepresence robot, resulting in the development of a Collaborative LLM-based Embodied Assistant Robot, named CLEAR Robot (Figure 2). A controlled experiment simulating a remote assembly task of one to two demonstrated that, compared to the standard telepresence robot, CLEAR Robot significantly improved efficiency, reduced cognitive load, facilitated more balanced collaboration, and improved the user experience. We also discuss the impact of language-driven implicit interactions in multi-user collaboration and provide insights for designing robot systems that support one-expert-multiple-workers remote guidance in the future.

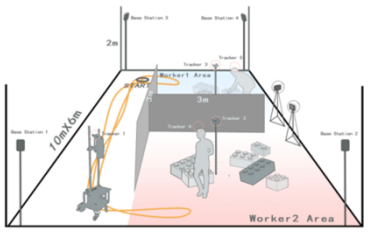

Figure 2 *The CLEAR Robot interface reduces the cognitive load of one-expert-multiple-worker remote assistance through following the navigation intents in the expert's natural*

Intelligent Training and Learning Systems with Multimodal Understanding and Feedback

Theme 1: Multimodal Tools for Information Gathering

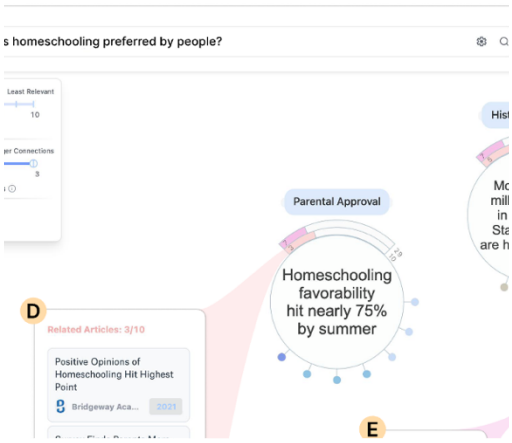

Figure 3 Compendia synthesize key themes and visualize them from online sources.

In the digital age, readers value quantitative journalism that is clear, concise, analytical, and human-centred. To understand complex topics, they often piece together scattered facts from multiple articles. Visual storytelling can transform fragmented information into clear, engaging narratives, yet its use with unstructured online articles remains largely unexplored. To fill this gap, we present Compendia, an automated system that analyzes online articles in response to a user's query and generates a coherent data story tailored to the user's informational needs. Compendia addresses key challenges of storytelling from unstructured text through two modules covering: Online Article Retrieval,

which gathers relevant articles; Data Fact Extraction, which identifies, validates, and refines quantitativefacts; Fact Organization, which clusters and merges related facts into coherent thematic groups; and Visualization, which transforms the organized facts into narratives with visualizations in an interactive scrollytelling interface. We evaluated Compendia through a quantitative analysis, confirming the accuracy in fact extraction and organization, and through two user studies with 16 participants, demonstrating its usability, effectiveness, and ability to produce engaging visual stories for open-ended queries.

Theme 2: Training Systems in Embodied Environments

Designing adaptive tutoring systems for software learning presents challenges in determining appropriate instructional modalities. To inform the design of such systems, we conducted an observational study of ten human teacher-student pairs (N=20), where experienced design software users taught novices two new graphic design software features through multi-step procedures. These lessons were limited to three communication channels (speech, visual annotations, and remote screen control) to mimic possible AI tutor modalities. We found that annotations complement speech with spatial precision and remote control complements it with spatial and temporal precision but both of them cause intrusion to learner agency. Teachers adaptively select modalities to balance the need for instruction progress with students' cognitive engagement and sense of digital territory ownership. Our results provide further support to the contiguity principles and the value of agency in learning, while suggesting precision-agency trade-off and digital territoriality as new design constraints for adaptive software guidance.

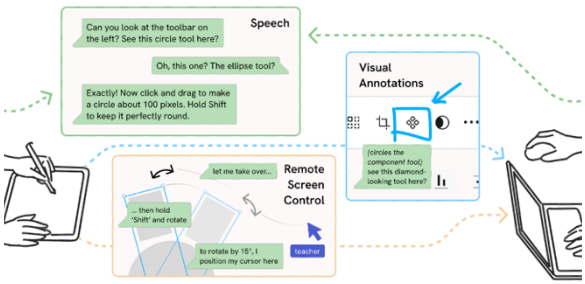

Figure 4 Our study looked how multiple modalities are used in concert in software instruction

Selected Publications and Outputs

Show It, Don’t Just Say It’: The Complementary Effects of Instruction Multimodality for Software Guidance. In submission.

Karthik Mahadevan, Blaine Lewis, Jiannan Li, Bilge Mutlu, Anthony Tang, Tovi Grossman. ImageInThat: Manipulating Images to Convey User Instructions to Robots. HRI 25.

Ruyi Li, Jingfei Guo, Xinyi Zhang, Xuji Zhang, Zeqing Li, Jiannan Li, Jiangtao Gong. Guiding Multiple Remote Users in Physical Tasks with LLM-powered Robotic Telepresence. IMWUT 25.

Compendia: Automated Visual Storytelling Generation from Online Article Collection. In submission.