Research Statement

LI Jiannan
School of Computing and Information Systems, Singapore Management University
Email: jiannanli@smu.edu.sg
19 (Day) 12 (Month) 2024 (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. uncertainty of AI output, instructions to follow) to human users. As a next step, these theories are applied to a range of possible scenarios using foundational models and other computational techniques. 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.

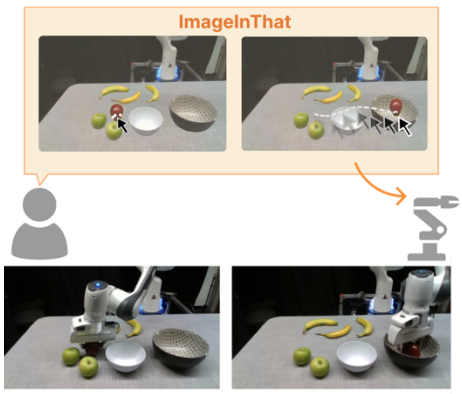

Figure 1. We introduce the direct manipulation of images as a paradigm for providing instructions to a robot. Depicted in the bottom left are a series of instructions that a user is giving the robot by manipulating the fruits. The top shows one trajectory of direct manipulation.

Research Areas

Collaborative Robots that Learn Human Intents through Multimodal Signals

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 1Figure 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.

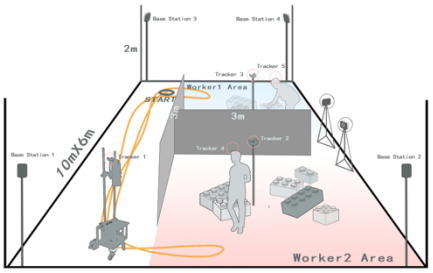

Figure 2 The CLEAR Robot interface reduces the cognitive load of one-expert-multiple-worker remote assistance through following the navigation intents in the expert's natural language and visualizing the expert's spoken instructions.

Intelligent Training and Learning Systems with Multimodal Understanding and Feedback

Theme 1: Multimodal Tools for Fostering Creativity

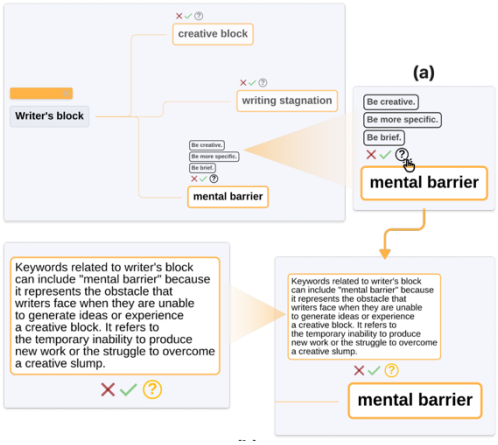

Figure 3 Polymind integrates LLM-powered ideation into visual diagramming using microtasks, while preserving user agency to facilitate iteration.

Prewriting is the process of generating and organizing ideas before a first draft. It consists of a combination of informal, iterative, and semi-structured strategies such as visual diagramming, which poses a challenge for collaborating with LLMs in a turn-taking conversational manner. Polymind is a visual diagramming tool that leverages large language models to support prewriting (Figure 3). The system features a parallel collaboration workflow in place of the turn-taking conversational interactions. It defines multiple "micro-tasks" to simulate group collaboration scenarios such as collaborative writing and

group brainstorming. Instead of repetitively prompting a chat-bot for various purposes, Polymind enables users to orchestrate multiple microtasks simultaneously. Users can configure and delegate customized micro-tasks, and manage their micro-tasks by specifying task requirements, toggling visibility and initiative. Our evaluation revealed that Polymind was able to quickly expand writing ideas and efficiently organize existing diagrams on a canvas.
My ongoing work along this line of research is exploring a human-AI workflow to support scientists and educators in creating comics to communicate scientific knowledge and concepts to students and the public.

Theme 2: Multimodal Tools for Learning Advanced Software Features

Software skills are critical for productivity but software tools are increasingly complex to use and to learn. AI-powered tutoring systems can provide users with step-by-step customized guides for unfamiliar software tasks, but these instructions do not necessarily promote effective learning due to learners' low cognitive engagement when the next action is immediately obvious to them. Through a study where participants learned to use a feature-rich design application following step-by-step guidance, we first show that visual step-by-step guidance led to better initial performance than text-based guidance, but they were equally insufficient for learning (Figure 4). We are conducting another study to investigate whether adding rationale for key steps to step-by-step guidance could improve learning outcome.

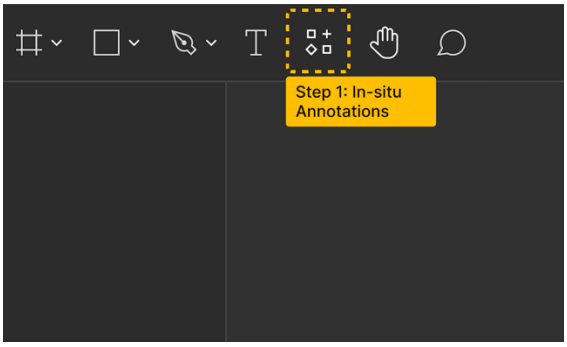

Figure 4 Step-by-step visual instructions show the next step to perform in the interface; while easy to follow, they do not necessarily lead to effective learning due to learners' low cognitive engagement when the next action is immediately obvious.

Selected Publications and Outputs

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. Under submission.

Qian Wan, **Jiannan Li**, Huanchen Wang, Zhicong Lu. Polymind: Microtask-enhanced Visual Diagramming with Large Language Models to Support Prewriting. Under submission.