Research Statement

LI Jiannan
School of Computing and Information Systems, Singapore Management University
Tel: (65) 68264834; Email: jiannanli@smu.edu.sg
24 (Day) 12 (Month) 2023 (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

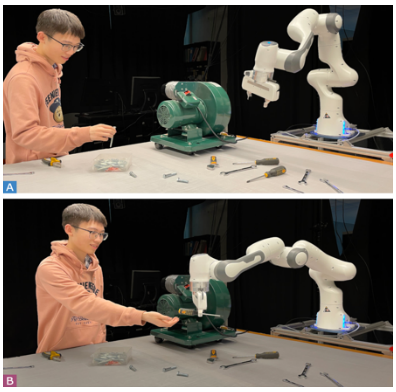


Figure 1 LLMs can reason about human intents and perform complementary actions. Here, the robot hands over the human a screwdriver (B) for the screw he just picked up (A).

Collaborative Robots that Learn Human Intents through Multimodal Signals

Theme 1: Human-Robot Collaboration based on Large Language Models

Designing for human-robot collaboration (HRC) is difficult due to the diversity and unpredictability of human behavior while task-specific policies are hard to generalize to arbitrary scenarios. My work proposes a novel approach for HRC that utilizes large-language models (LLMs) to synthesize information about the task, environment, and human partner to generate robot behaviors to facilitate the collaborative task. The new system, LingoBot, demonstrates this approach on a tabletop robot arm within a pick-and-place task context (Figure 1). LingoBot actively monitors the environment, user's actions, and dialogue, and autonomously determines appropriate actions to assist in the collaboration. Actions can include proactively passing nearby objects to the human or reactively assisting by mirroring the human's actions. In an evaluation, participants performed collaborative sorting across a range of task conditions with LingoBot. The results suggest that LLMs can generate complementary actions to support the human-robot partnership across a diverse range of tasks and contexts. In future work, I plan to enhance this collaboration workflow by enabling the robot to learn human users' preferences by observing their corrective actions (e.g. undo an action the robot has just performed).

Theme 2: Camera Robots

Videos can be used to teach critical hands-on skills for vocational training and everyday applications. Professional productions of such content often employ dedicated camerapersons. They follow the instructors and film the demonstrations of the skills from varying distances and perspectives to highlight important details and to create visual interest. However, most instructors making these tutorials do not have access to the help of a camera crew and must work in the constraints of fixed camera setups, which often fall short of flexibility. Robots can manipulate cameras in place of camera operators but lack the contextual intelligence of the latter to make sensible filming decisions. As an application of the multimodal human intent approach, part of my research focuses on building human-robot synergy, in which camera robots respond in real-time to instructors to create more informative and engaging video tutorials.

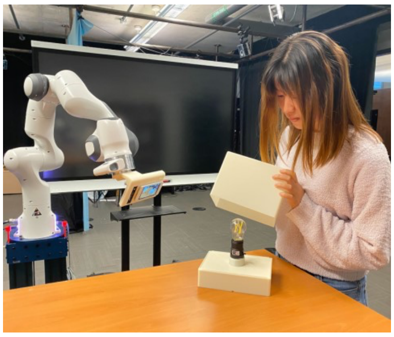
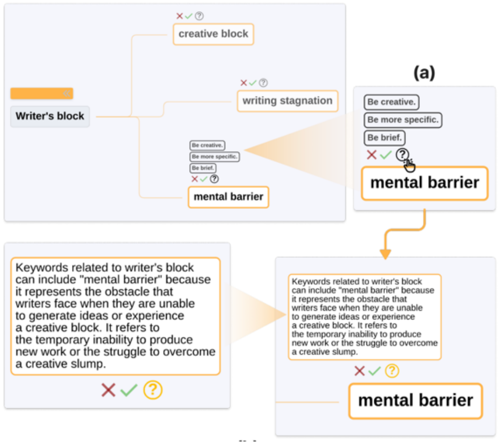


Figure 2 An instructor recording a tutorial on making a 3D-printed lamp with Stargazer.

Stargazer explores a synergetic workflow in which a highly articulated camera robot follows and records the instructor's actions mostly autonomously (Figure 2). As the instructor goes about showing the skills, they can leverage gestures and speech that they already use to address the *audience* to adjust the robot's behaviors without disrupting instruction delivery. For instance, the instructor can have the robot change its view to look at each of the tools they will be using during the tutorial—pointing at the tools one after another to have the camera pan to show each of them. They can also say to the audience, "If you look at how I install this from the top". The robot will parse the speech with a large language model and respond by framing the instructor's action with a high camera angle, giving the audience a better view. The robot uses optimization-based real-time motion planning to react promptly while still following smooth camera trajectories. Working together with Stargazer, instructors were able to create video tutorials for a wide range of skills, from VR equipment setup to interactive sculpture making, and were satisfied with the videos' quality.

Intelligent Training and Learning Systems with Multimodal Understanding and Feedback

Theme 1: Multimodal Tools for Fostering Creativity and Learning



Keywords related to writer's block can include "mental barrier" because it represents the obstacle that writers face when they are unable to generate ideas or experience a creative block. It refers to the temporary inability to produce new work or the struggle to overcome a creative slump.

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: Training Systems in Embodied Environments

Educational psychologists have noted the perspective effect, i.e. first-person perspective video tutorials promote physical skill learning because of the reduced discrepancy between the demonstrations and the learner's perception of space and actions. Through an empirical study, I investigated whether the immersion and viewer agency afforded by VR videos can further close this gap between demonstrations and perception, and consequently enhance learning. The study compared people's learning outcome and perceptions with first-person VR videos (Figure 4), first-person 2D videos, and third-person 2D videos using everyday motor tasks. Results highlighted the benefits of first-person VR videos in accelerating learning and reducing learners' cognitive load.

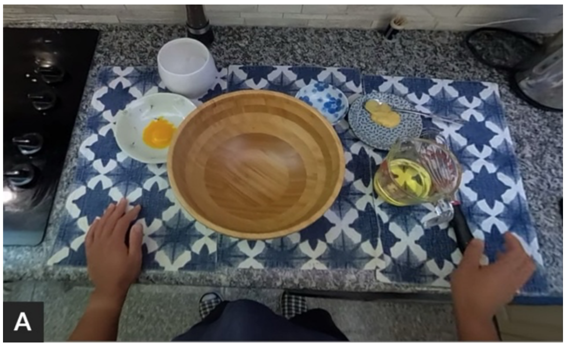


Figure 4 First-person VR videos can benefit the learning of motor skills, such as cooking

Building on this thread of research, one of my current projects is developing an augmented reality assistance system that choose appropriate modalities (text vs. image vs. anchored visualizations) based on users' actions.

Selected Publications and Outputs

Jiannan Li, Maurício Sousa, Karthik Mahadevan, Bryan Wang, Paula Akemi Aoyagui, Nicole Yu, Angela Yang, Ravin Balakrishnan, Anthony Tang, Tovi Grossman. Stargazer: An Interactive Camera Robot for Capturing How-To Videos Based on Subtle Instructor Cues. CHI '23.

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

Kevin Huang, **Jiannan Li**, Maurício Sousa, Tovi Grossman. immersivePOV: Film-ing How-To Videos with a Head-Mounted 360° Action Camera. CHI ’22.