Do, Look, Point, or Tell: Understanding the Cognitive power of Human-Robot Interaction under Dual-Task Conditions

Gyanig Kumar, Raktim Bijoypuri, Shivendra Agarwal, Bradley Hayes

Response to Prof. Brad's Feedback

The secondary pipeline would take coming 1-2 weeks to completely integrate. We plan to leverage existing components from ongoing projects, including the gaze module, camera system, and teleoperation mode, to accelerate development. We have started working on the integration of the gaze, pointing and verbal modules already. While ambitious, we are committed to building this as closely as possible to our proposal. Raktim and I will check in during office hours to monitor progress and ensure a meaningful project outcome by course end.

Response to Shivendra's Feedback

On Distractor Task (1st Point)

We agree that clicking on the laptop for the distractor task will suffice as an alternative to a foot pedal.

On Secondary Task (2nd Point)

The secondary task, of Lego building require users to follow a fixed number of steps (10 per se) without able to mix the sequence of building the final structure. We would mix and match colour to generate triangular pyramid structure for the user to complete. The instruction would be provided to the user via paper/laptop screen displayed for each step, the user has the option to select Yes and move to the next block and assemble them according to image provided. The reason of our approach of step by step would eliminate the need for find different sub tasks within the building the structure enabling the user to utilize the robot to the maximum extent possible.

On VLM Selection and Assumptions (3rd Point)

Initially, we will use a hardcoded approach for object pickup (Lego blocks) across modalities: distinct objects, well-spaced for gaze/pointing, and specific verbal commands. This enables rapid pipeline prototyping. Subsequently, we will integrate the Qwen3-VL-4B-Instruct model, runnable locally, to demonstrate agentic capabilities in robot tasks without API costs (e.g., GPT/Gemini credits).

Step-by-Step Build Plan

- **System Foundations**: Depth camera for <5cm object position estimation; robot alignment for grasping (testing GraspNet/AO-Grasp); joystick teleoperation via relaxed-IK; ROS wrapper with Tobii eyetracker.
- **a) Initial Modalities (Gaze & Verbal)**: Implement intent inference and robot action execution per proposal diagram. Use existing gaze module for object detection; start verbal with exact keywords (no VLM) for pipeline testing. *(Target: Early October)*
- **b) Add Remaining Modalities**: Integrate teleoperation (existing Sawyer joystick with joint velocity control) and pointing (triangulate finger direction via depth camera; assess error in complex scenarios). Maintain well-defined setup for reliable grasping.
- **c) Unit Testing & Partial Pipeline**: Validate all modalities and establish experiment flow. *(Current Goal: End of October)*
- **d) VLM Integration**: Test Qwen3-VL for multi-modal intent inference.
- **e) Refinement & Pilot Study**: Address issues for consistent modality timing; conduct pilot with 2-5 participants. *(Goal: End of November)*

Problem Statement

Understanding human intent has been an integral research question, when it comes to human robot interactions (HRI). The non-verbal communication or modality like gaze has been an important indicator of intent, showing the internal reward system of a human brain directly [Saran, Akansksha et al. 2020]. Verbal communications to a robot via LLM [Wang, Chao, et al. 2024] or VLM [Liu, Huihan, et al. 2025] shows the use-cases of explicit instructions used for robot tasks. While current HRI systems can process individual modalities like gaze or language efficiently, they are fundamentally "state-unaware.", which means that they often lack awareness of the user's real-time cognitive load, treating all interactions as if they occur under ideal conditions. In the daily life scenarios, if we are designing a social robot, who is engaging with a human to aid to the request like bring him a water bottle from the red table which is two brown tables far away while speaking over his phone, he can choose to share his/her intent via various ways, which the robot needs to interpret and aid the human. Human might say, "...yeah, grab me the bottle from the red table over there," while simultaneously glancing at the intended object. Or to fulfill this, the human might employ a combination of verbal commands and deictic gestures like pointing. Here, the verbal instruction could be ambiguous ("the bottle," "over there"), but the non-verbal like pointing or gaze cue might be less ambiguous. But before we jump into modelling a framework where the robot infer/adapts to the user's intended target and various actions intelligently by fusing information from multiple communication channels, we have not learned anything about preferences on modalities, that comes natural to humans. A foundational understanding of the human side of the equation seems to be lacking. While the prior work has validated the technical efficacy of individual modalities like gaze or natural language, there is a lack of empirical data on the comparative cognitive cost of using these modalities, especially in realistic scenarios where users are multitasking. We lack a clear empirical picture of the inherent trade-offs between modalities, potential correlation and examine the needs of different modalities. We aim to explore the different modalities like

teleoperation (no intent situation: baseline), eye gaze, hand pointing and verbal modalities. Our work aims to complete the missing piece of cognitive understanding of humans, who utilize the different modalities as per their need.

We are pursuing this research direction because we want to bridge that gap between robotics and human factors. We design our experiment, where the user is engaged in a primary task i.e. solving random arithmetic problems and a secondary task to build a Lego structure based on a given set of instructions. The human cannot reach for the Lego blocks, so he/she is given a chance to use each modality to provide instructions to the robot for picking up an intended object. Through the experiment, we will answer the following foundational questions:

1. How does the choice of interaction modality influence the user's perceived cognitive workload?
2. What are the objective trade-offs in task performance (speed, efficiency, and error rate) across the different modalities?
3. Which modalities do users qualitatively prefer when engaged in a multitasking scenario, and what reasons underlie these preferences?

Related Works

This proposal is grounded in established research on communication modalities in HRI. Gaze has been identified as a powerful, low-effort channel for conveying intent, particularly in learning from demonstration settings [Saran, Akanksha et al., 2020; Huang, Sandy H., et al.]. [Koppula, et al. 2015] focused on how a robot can proactively assist a human by not just observing where they are looking now, but by learning to predict their future gaze and body movements. The combination of gaze and joystick control has also proven effective in shared autonomy tasks [Aronson, Reuben M., and Henny Admoni, 2022]. On the other hand, the expressiveness of natural language, unlocked by LLMs and VLMs, provides a direct and explicit means of communication [Wang, Chao, et al. 2024; Liu, Huihan, et al. 2025]. Another paper from Google Robotics [Ahn, Michael, et al 2022] introduces a powerful "SayCan" system combining a **Large Language Model (LLM)** with a robot's learned skills (**affordances**). The LLM proposes plausible next steps to fulfill a command, but the robot's own model of what it can do in its environment is used to select the most feasible action. While these works validate the power of each modality, they typically focus on the technical implementation rather than the human cost of using them. Assuming that humans will intuitively embrace robots for daily tasks overlooks the significant learning curves and social adjustments required for adoption.

For instance, research [Gombolay et al. 2024] shows how to account for “nonspherical” aspects of human and help people express their priorities for team coordination. [Šabanović, Selma, et al. 2010] argues why your human-centric study is not just useful, but essential for designing the next generation of robots. This research aims to fill that gap by providing a systematic, human-centric evaluation of these modalities, creating an empirical foundation for designing future adaptive HRI systems. This work draws on course concepts from Experimental Design, Gaze in Robotics, Social Robots, in all essence of algorithmic human robot interaction. It would a fun demonstration of different modalities with Sawyer.

System Plan

Robot: The experimental testbed will be built in ROS, with deployment on the physical Sawyer collaborative robot arm.

Algorithms: The system will be architected as a robust experimental platform capable of switching between four distinct modality-handling modules:

Perception: A YOLOv8 model will continuously detect and track Lego blocks in the workspace.

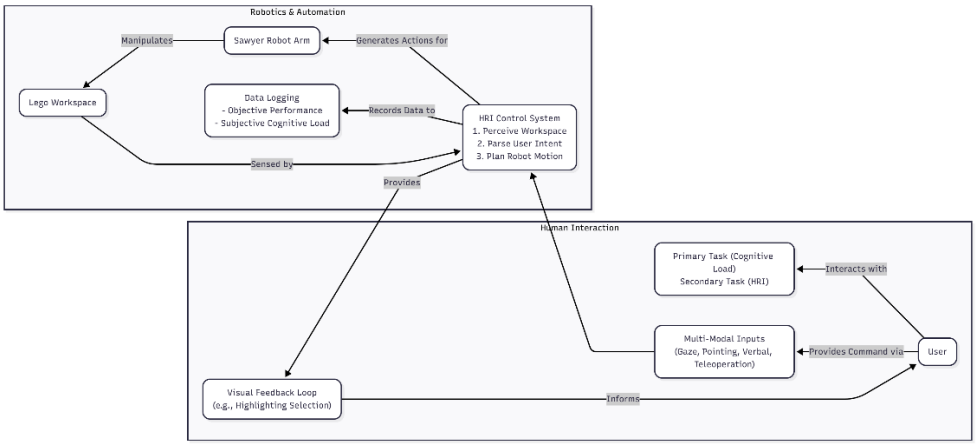
Teleoperation Module: A standard joystick will provide end-effector control.

Gaze Module: A Tobii Glasses 2 eye-tracker will trigger a "selection" based on a sustained dwell time on a specific Lego block.

Pointing Module: An Intel RealSense camera and MediaPipe will triangulate the 3D position of a pointing gesture.

Verbal Module: A microphone and local LLM will parse simple commands (e.g., "I need the blue 2x2 brick on the left top of the table").

The user would be prompted randomly different modalities to test while doing the task.



Metrics:

Primary Metric (Cognitive Load): The NASA Task Load Index & System usability Score (NASA-TLX & SUS), a validated survey, will be administered after each condition to measure perceived workload of human and usability of the system.

Objective Performance Metrics: Task Completion Time on both tasks, Primary Task Accuracy : percentage of correct True/False answers on the arithmetic problems, Action Efficiency : time per Lego piece, and Error Rate : incorrect or dropped pieces.

Qualitative Metrics: A post-experiment survey will capture user rankings of the modalities, subjective preferences, and general feedback.

Evaluation Plan

This would be a non-IRB-approved evaluation, basing it on pilot study with few participants(friends).

Experimental Design: We will use a **within-subjects design**, where every participant completes the task under all four modality conditions. The order of the conditions will be random or counterbalanced across participants to mitigate learning effects.

Scenario & Tasks:

- **Primary Task (Cognitive Load):** To induce a consistent but lightweight cognitive load, the primary task will be a **self-paced arithmetic verification task**. Participants will be shown a continuous stream of simple arithmetic equations (e.g., 8 + 5 = 13 or 9- 2 = 6) on a monitor. A new equation will be available every 10 seconds with time to solve the question and extra time to engage with robot before new question appears. For time exceeding, a penalty to the point would be considered. To avoid interference with the verbal HRI modality, participants will respond to each problem by judging it 'True' or 'False' using a two-button foot pedal.
- **Secondary Task (HRI):** Participants will follow a visual instruction set to build a small Lego model with 10 pieces. Having more pieces would mean that the participant is engaged in the secondary task as well. They will use the assigned modality for that condition to instruct the Sawyer robot to pick a specific Lego block from a cluttered table and place it near him for assembly.

We will use an ANOVA test to determine if there are significant differences in the NASA-TLX scores and objective performance metrics across the four conditions.

Risks & Mitigation

- **System Integration/Setup problems** - The experimental testbed must be robust. We would measure the "time-to-selection" for each module under ideal conditions and ensure they are roughly equivalent, so that different modalities are on same system latency. Each ROS node will be unit-tested extensively. We will run pilot studies to debug the full system and ensure smooth transitions between conditions.
- **Problems with parsing the intent** - Gaze tracking, gesture recognition, and speech-to-text are imperfect. The system will provide clear visual feedback to the user (e.g., highlighting the "selected" block) so they can immediately correct any parsing errors. Upon a potential selection from any modality, the corresponding object's bounding box will be highlighted in green on the user-facing monitor for defined time slots like 2s, allowing the user to confirm or cancel the action.
- **Problems with verbal parsing** – Testing LLM outputs for parsing the intended object could be difficult with open-source models. Using Google gemini student subscription, it is still a difficult problem. Backup is using more robust, grammar-based system or a simpler intent-classification model e.g. using OpenAI Whisper for transcription followed by keyword spotting.

1st stage (October to 1st November) – we plan to test the two modalities i.e. **gaze and verbal** as the starting basis of developing the pipeline. (Gaze - Gyanig, Verbal – Raktim)

2nd stage (2nd November to 20th November) – we plan to add the two more modalities for testing. (Teleoperation – Gyanig, Pointing – Raktim)

3rd stage (21st November to 4th December) – Performing some pilot studies, and writing the paper (Gyanig & Raktim together)

Presentation and submission – 5th December

Resources needed

Compute: A dedicated workstation with an NVIDIA RTX 3080 or higher GPU.

Hardware: Access to the lab's Sawyer robot arm, an Intel RealSense D435 camera, a Tobii Glasses 2 eye-tracking headset, an Xbox joystick, and a two-button USB foot pedal.

Participants: Asking friends ~5-10 participants for the pilot study. Increase to more in future

Other: A supply of standardized Lego kits for the building task.

References

1. Saran, Akanksha, et al. "Understanding teacher gaze patterns for robot learning." Conference on Robot Learning. PMLR, 2020.
2. Liu, Huihan, et al. "Casper: Inferring Diverse Intents for Assistive Teleoperation with Vision Language Models." arXiv preprint arXiv:2506.14727 (2025).
3. Wang, Chao, et al. "Lami: Large language models for multi-modal human-robot interaction." Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 2024.
4. Koppula, Hema S., Ashesh Jain, and Ashutosh Saxena. "Anticipatory planning for human-robot teams." *Experimental Robotics: The 14th International Symposium on Experimental Robotics*. Cham: Springer International Publishing, 2015.
5. Ahn, Michael, et al. "Do as i can, not as i say: Grounding language in robotic affordances." *arXiv preprint arXiv:2204.01691* (2022).
6. Gombolay, Matthew. "Human-Robot Alignment through Interactivity and Interpretability: Don't Assume a "Spherical Human"." *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 2024.
7. Šabanović, Selma. "Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design." *International Journal of Social Robotics* 2.4 (2010): 439-450.