2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Integrating Multimodal Communication and Comprehension Evaluation during Human-Robot Collaboration for Increased Reliability of Foundation Model-based Task Planning Systems

Eden Martin[1], Shoichi Hasegawa[2], Jorge Solis[3], Benoit Macq[1], Renaud Ronsse[1], Gustavo Alfonso Garcia Ricardez[2], Lotfi El Hafi[2,*], and Tadahiro Taniguchi[2,4]

***Abstract*— Foundation models provide the adaptability needed in robotics but often require explicit tasks or human verification due to potential unreliability in their responses, complicating human-robot collaboration (HRC). To enhance the reliability of such task-planning systems, we propose 1) an adaptive task-planning system for HRC that reliably performs non-predefined tasks implicitly instructed through HRC, and 2) an integrated system combining multimodal large language model (LLM)-based task planning with multimodal communication of human intention to increase the HRC success rate and comfort. The proposed system integrates GPT-4V for adaptive task planning and comprehension evaluation during HRC with multimodal communication of human intention through speech and deictic gestures. Four pick-and-place tasks of gradually increasing difficulty were used in three experiments, each evaluating a key aspect of the proposed system: task planning, comprehension evaluation, and multimodal communication. The quantitative results show that the proposed system can interpret implicitly instructed tabletop pick-and-place tasks through HRC, providing the next object to pick and the correct position to place it, achieving a mean success rate of 0.80. Additionally, the system can evaluate its comprehension of three of the four tasks with an average precision of 0.87. The qualitative results show that multimodal communication not only significantly enhances the success rate but also the feelings of trust and control, willingness to use again, and sense of collaboration during HRC.**

I. Introduction

Humans and robots often work separately on independent tasks to maximize efficiency and safety. However, collaboration can enhance certain tasks by establishing a shared workspace and mutual understanding between agents, which is suitable for real-life tasks in various environments. To achieve this, robots need the ability to understand semantic meanings, enabling them to adapt to human teammates, context, and diverse task settings.

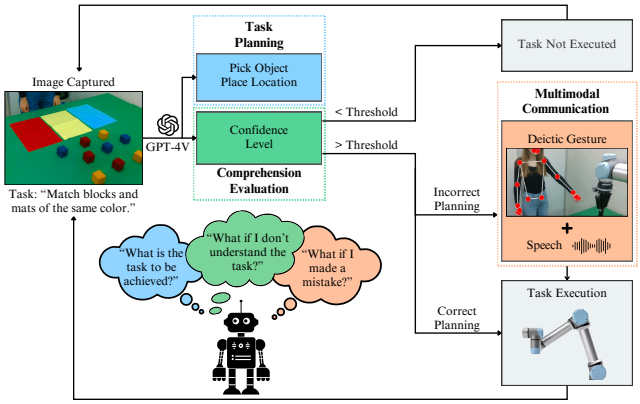


Fig. 1. The proposed system enhances the reliability of foundation model-based task planning in HRC (blue area) through comprehension evaluation using an inferred confidence level by the robot agent (green area) and multimodal communication using deictic gestures and speech instructions by the human agent (orange area).

Conventional programming requires models to be trained on specific and extensive data, which limits flexibility in adapting to different environments, tasks, and objects [1], [2]. Recent research has explored and adopted the use of large pre-trained models for adaptability, eliminating the need for additional training for specific tasks. In this regard, recent breakthroughs with transformers [3] and subsequent foundation models, such as large language models (LLMs) like OpenAI GPT [4], have shown promising potential for solving complex robotics tasks [5], [6]. These models have also allowed adaptability and zero-shot learning improvements in robotics [7], [8], enabling robots to acquire human-like reasoning capabilities and understand unspecified tasks.

However, when responding to prompts, multimodal LLMs do not provide information about the reliability of their answers. It is, therefore, essential for the human agent to know whether the robot is confident in its answers. Additionally, in the case of unwanted behavior by the robot, it is necessary to maintain human control to ensure correct task achievement, for instance, by communicating intentions.

This leads to the following research questions:

Q1: How can we enable an adaptive robotic task-planning system for human-robot collaboration (HRC)?

Q2: Can multimodal LLMs evaluate their own comprehen-

This work was supported by the Japan Science and Technology Agency (JST), Moonshot Research & Development Program, Grant Number JPMJMS2011.

[1]Eden Martin, Benoit Macq, and Renaud Ronsse are with Université catholique de Louvain (UCLouvain); 1 Place de l'Université, Louvain-la-Neuve 1348, Belgium. eden.martin@student.uclouvain.be, {benoit.macq, renaud.ronsse}@uclouvain.be

[2]Shoichi Hasegawa, Lotfi El Hafi, Gustavo Alfonso Garcia Ricardez, and Tadahiro Taniguchi are with Ritsumeikan University; 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan. {hasegawa.shoichi, lotfi.elhafi, garcia-g, taniguchi}@em.ci.ritsumei.ac.jp

[3]Jorge Solis is with Karlstad University; 2 Universitetsgatan, Karlstad 651 88, Sweden. jorge.solis@kau.se

[4]Tadahiro Taniguchi is with Kyoto University; Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan. taniguchi@i.kyoto-u.ac.jp

*Corresponding author.

sion of a collaborative task?

Q3: How can communicating intentions enhance the task success rate and comfort (*e.g.*, feeling of trust or willingness to use again)?

Regarding Q1, previous research demonstrates that multimodal LLMs can interpret tasks from images by breaking down complex, high-level instructions into smaller, actionable steps. For instance, a broad task like "put the red block in the blue box" is divided into specific actions, such as "grab the red block" [9], [10], [11]. However, these methods use pre-defined low-level steps, reducing adaptability capabilities. Another approach is to ask a multimodal LLM to learn a task through a video [12], but this method lacks real-time capabilities and requires human supervision.

Regarding Q2, a common way to assess confidence in the answers of models is to repeat the same request and compare the answers [13], [14]. Although these previous studies show that the models understand the meaning of confidence, their proposed methods can lead to a time-consuming approach if used with the analysis of images.

Regarding Q3, intention recognition can be achieved either by gesture recognition or voice recognition separately, but the simultaneous use of both communication methods enables robots to perform a wider variety of tasks [15]. For example, if the instruction "pick this" is given, what does the "this" in the sentence refer to? One type of gesture, deictic movement (pointing movement), is natural and familiar to humans and helps disambiguate speech communication [15], [16], [17].

Therefore, we propose a multimodal task-planning system for HRC, outlined in Fig. 1. In particular, the contribution of this paper is two-fold:

1) An adaptive task-planning system for HRC that reliably performs non-predefined tasks implicitly instructed through HRC.
2) An integrated system combining multimodal LLM-based task planning with multimodal communication of human intention to increase the HRC success rate and comfort.

Adaptive task planning is achieved using a multimodal LLM, GPT-4V, which provides the object to pick and the placement position for the next step of the task. Comprehension evaluation is done using a numerical confidence level also inferred by GPT-4V. Error correction is performed by communicating human intention using speech and deictic gestures during HRC.

Four pick-and-place tasks are used for the evaluation of the proposed system: color matching, object packing, and two object sorting tasks. In these experiments, the inferred confidence level is used as a threshold below which the robot is not allowed to act. Speech and deictic movements from the human instruct the robot when automated task planning fails to meet expected results.

The remainder of this paper is organized as follows. Section II introduces the related work. Section III describes the proposed system. Section IV presents the experiments and their results. Section V discusses the results. Finally, Section VI concludes with avenues for future work.

II. Related Works

A. Adaptive Task Planning with Foundation Models

Foundation models are now being used across a wide range of robotic applications as task planners [9], [10], [11], [12], [2], [18], [19] and interpreters of speech and gestures [15]. In particular, multimodal LLMs are commonly employed for task decomposition, where they break down high-level tasks into multiple low-level tasks [9], [10], [11], as well as for one-shot learning [12]. While these methods show promising results, they often rely on a pre-defined set of tasks or human supervision.

In this paper, we utilize a multimodal LLM to enable the robot to autonomously detect real-time, non-predefined tasks from images, facilitating dynamic HRC without requiring reprogramming or human supervision.

B. Comprehension Evaluation in Foundation Models

When responding to prompts, multimodal LLMs do not inherently provide information regarding the reliability of their answers. Research has been conducted to measure this reliability [13], [14], particularly in medical diagnosis. Kotelanski *et al.* [13] compared three methods: intrinsic confidence, self-consistency agreement frequency, and chain-of-thought response length. Intrinsic confidence involves the model rating its confidence on a 0-100 scale, with higher confidence levels assumed to indicate greater reliability. Self-consistency agreement frequency involves running the process 11 times and selecting the most common result, while chain-of-thought response length assumes that correct answers are generally longer. Although self-consistency agreement frequency proved to be the most accurate, it is also the most time-consuming, making it impractical for real-time applications. Li *et al.* [14] explore whether LLMs can assess their confidence in both deterministic (single solution) and open-ended tasks (multiple solutions). They propose two methods: consistency-checking, which involves running the request multiple times, and self-evaluation, using an if-or-else prompt to which models are asked to maintain or update their answers based on their confidence. Their experiments show that if-or-else prompts effectively evaluate the reliability of LLM responses, indicating that models have an understanding of confidence.

In this paper, based on the intrinsic confidence method, the robot's comprehension of the current task is evaluated by requesting a confidence level, defined as a numerical value.

C. Multimodal Communication of Intention in HRC

Effective communication is essential in all interactions [20] and must be intuitive for success and ease of use. Intention recognition is typically achieved through speech and gestures [21], [22], [23], [24], [25], [26]. In speech-gesture HRC systems, human agent intention, such as a request to pick up an object, is typically conveyed through spoken language, while attention, such as identifying a specific object, is indicated via gestures [17], [16]. The GIRAF system combines speech and gestures to enhance communication and success rates [15]. Deictic gestures,

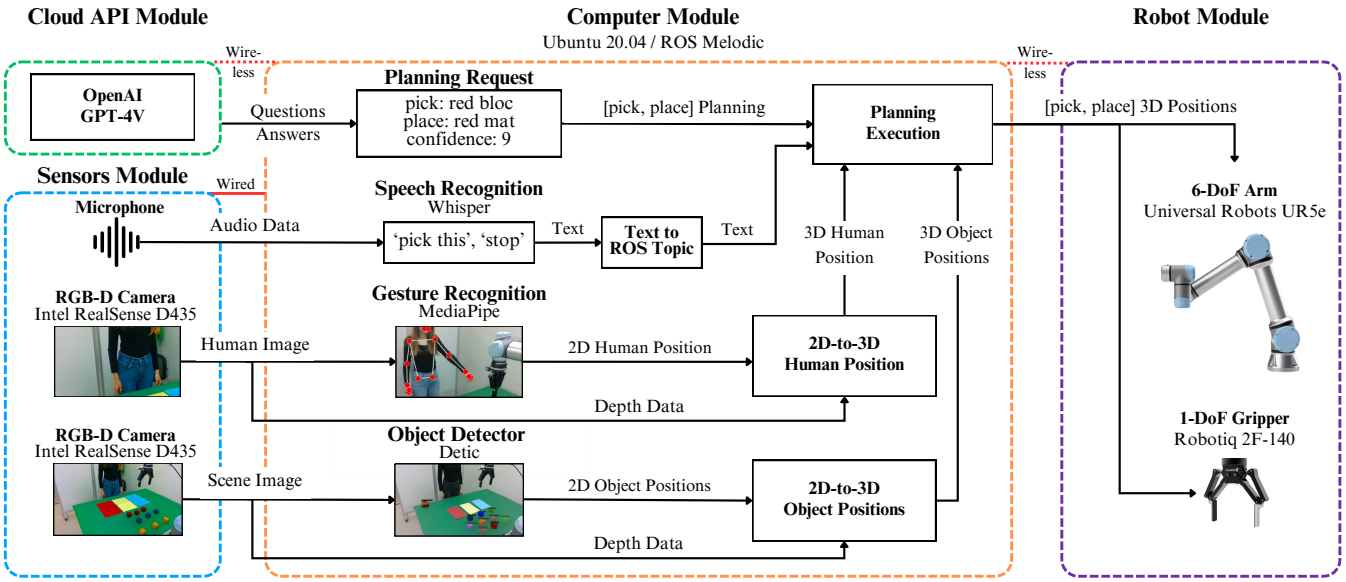


Fig. 2. Implementation of the proposed system. The colored areas represent the four modules of the proposed system: the cloud API, the sensors, the computer, and the robot. The black arrows indicate the flow of information, while the red lines denote the types of connections between the modules.

like pointing, are used to clarify exophoric terms such as "this", "that", or "here", thereby simplifying speech communication. In vision-based approaches, human pose estimation models can detect deictic gestures by analyzing relationships between body joints. Three main models for estimating pointing direction have been developed by [27]: elbow-wrist, head-wrist, and shoulder-wrist. The elbow-wrist model loses accuracy in large environments, while the head-wrist model faces occlusion issues during pose estimation, such as interference from helmets [28]. By calculating the cosine similarity between the shoulder-elbow and elbow-wrist vectors, their system determines whether the movement is deictic.

In this paper, deictic gestures based on the shoulder-wrist model are integrated into the task planning system but only to instruct the robot when it makes an incorrect decision.

III. Proposed System

In this study, we propose a system to increase the reliability of task planning in HRC based on foundation models, particularly multimodal LLMs. An overview of the proposed system is shown in Fig. 2. The system integrates three key aspects: task planning, comprehension evaluation, and multimodal communication.

A. Task Planning

We propose using a multimodal LLM to achieve adaptability in task planning with respect to the task and context. We use GPT-4V because it effectively handles a variety of inputs and benefits from a large and diverse training dataset that includes both text and images. The term implicit will be used to qualify tasks that should be inferred solely by analyzing the content of an image, as opposed to providing an explicit task description as input to the model. The problem of the proposed automated task planning system is defined as follows: given only an image of the workspace, the robot must determine which object to pick and where to place it.

We use Detic [29] to detect available objects, given an image of the scene captured by a scene camera recording the workspace. The labels (*i.e.*, type or category) of the objects to detect are provided by the human agent. The positions of the objects in space are computed by transforming the 2D bounding boxes into 3D positions using the depth information obtained from the camera. The final output is a set $O = \{(c_1, x_1, y_1, z_1), ..., (c_i, x_i, y_i, z_i), ...\}$, where c_i represents the class label, and (x_i, y_i, z_i) represents the 3D position of object i relative to the robot's base.

B. Comprehension Evaluation

When using multimodal LLMs for task planning, it is essential for the proposed system to quantify its comprehension of the task and the reliability of its responses. We express the robot's confidence in its task comprehension as a confidence level computed by GPT-4V. The input is a fixed prompt along with an image of the workspace, and the output is a coefficient ranging from 0 (no confidence) to 10 (maximum confidence), indicating the system's self-evaluation of its comprehension of the task.

If this quantification of comprehension is correctly linked to task planning performance, incorrect planning can be detected and avoided. Once computed, we set a threshold below which the robot is not allowed to act if the confidence level is too low.

C. Multimodal Communication

The proposed system allows the human's speech and gestures to interrupt the automated task planning in case of failure during HRC (*e.g.*, placing an object in the wrong position). This is necessary for the human agent to maintain control over the robot's actions and has the advantage of

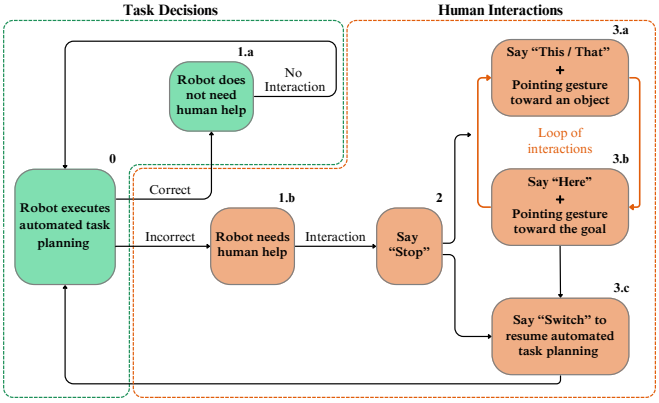
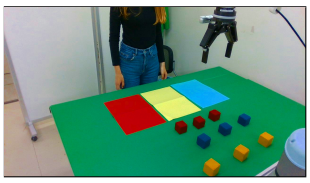


Fig. 3. Pipeline of the multimodal communication. GPT-4V handles task planning. If correct, the robot works autonomously. If not, human intervention is required. The human agent says "stop" and points to the correct object or area. "Here" or "there" are used to place an object, and "this" or "that" to pick one. Saying "switch" restarts the planning after human intervention. The human agent can continue directing the robot for sequential tasks.



(a) Task 1: Matching blocks and mats of the same color.

(b) Task 2: Packing either cans or bottles into the box.

(c) Task 3: Sorting bottles on the yellow mat and cans on the red mat.

(d) Task 4: Sorting bottles on the yellow mat, cans on the red mat, and leaving the blue mat empty.

Fig. 4. Overview of the four tasks, implicitly instructed through HRC, used to evaluate the performance of the proposed system.

being both rapid and intuitive. The problem of multimodal instruction communication is defined as follows: given a speech instruction S, for intention recognition, or a deictic gesture instruction D, for attention recognition, the robot must react according to these instructions with a human-controlled task planning T.

We use Whisper [30] for speech recognition. The audio captured by the microphone is continuously recorded and analyzed in frames of three seconds each. The text is then analyzed to determine if keywords such as "stop", "switch", "here", "there", "that", and "this" are present in the speech instructions.

The deictic gesture recognition process consists of two parts: deictic movement detection and pointed object detection. Deictic movement detection ensures that only pointing movements are considered, using keypoint positioning through MediaPipe [31]. This involves computing the cosine similarity between the vectors $\vec{v}_{\text{shoulder-elbow}}$ and $\vec{v}_{\text{elbow-wrist}}$. If the similarity is high, the gesture is classified as deictic. Similarly, pointed object detection uses the cosine similarity between the vectors $\vec{v}_{\text{shoulder-wrist}}$ and $\vec{v}_{\text{shoulder-object}}$. The object with the highest cosine similarity is selected as the pointed object:

$$\text{Obj.} = \underset{o}{\text{argmax}} \left(\text{cos. sim.}(\vec{v}_{\text{shoulder-wrist}}, \vec{v}_{\text{shoulder-object}})\right), \quad (1)$$

where o represents each object, and the argmax selects the object with the highest similarity. The complete multimodal communication pipeline is shown in Fig. 3.

D. System Integration

Fig. 2 describes the implementation of the integrated system. The system is divided into four modules: 1) the cloud API module that enables requests and responses using GPT-4V, 2) the sensors module that includes the Intel RealSense D435 RGB-D cameras and a microphone, 3) the computer module that executes the system within a containerized software development environment [32] based on ROS, and 4) the robot module consisting of a Universal Robotics UR5e arm and a Robotiq 2F-140 gripper. The computer is connected to the cameras and microphone through a wired connection, and to the cloud API, robotic arm, and gripper via a wireless connection.

The RGB images are used for human detection using MediaPipe and object detection using Detic. Their 3D positions are computed using the depth images. The proposed system computes the trajectory of the robot and the gripper based on the 3D positions of the human and objects captured by the cameras, speech recognition performed by Whisper via the microphone, and the instructions inferred by GPT-4V via the cloud API.

IV. Experiments

We designed four tabletop pick-and-place tasks to evaluate the planning performance of the proposed system in HRC, as shown in Fig. 4:

- T1: Matching blocks and mats of the same color.
- T2: Packing either cans or bottles into the box.
- T3: Sorting bottles on the yellow mat and cans on the red mat.
- T4: Sorting bottles on the yellow mat, cans on the red mat, and leaving the blue mat empty.

Each task is implicitly instructed through HRC, *i.e.*, solely inferred by the robot from the camera images without an explicit task description provided as input to the planning model. We also devised three experiments to evaluate each key aspect of the integrated system: task planning, comprehension evaluation, and multimodal communication.

A. Task Planning (Exp. 1)

This experiment aims to evaluate whether GPT-4V can detect an implicit task from an image and provide the correct next object to pick and position to place it. Two types of success criteria are assessed:

TABLE I

MEAN PLANNING SUCCESS AT EACH STEP FOR EACH TASK (EXP. 1)

Task	S0	S1	S2	S3	S4	S5	S6	S7	S8
T1	0.96	1	1	1	0.93	0.96	0.96	1	1
T2	0.56	1	0.90	0.73	-	-	-	-	-
T3	0.82	0.93	0.90	0.90	0.96	0.93	0.96	1	-
T4	0.30	0.50	0.47	0.50	0.37	0.27	0.70	0.77	-

TABLE II

WHOLE SYSTEM EXECUTION SUCCESS FOR EACH TASK (EXP. 1)

Task	Correct Compl.	Correct Place.	Time	GPT-4V Errors	Detic Errors	Other
T1	0.60	0.95	6.46	0	0.40	0
T2	0.80	0.98	3.27	0.10	0	0.10
T3	0.50	0.86	5.09	0.88	0	0.25

1) Planning success: Measures the effectiveness of GPT-4V in task planning. Success is determined by the correct alignment between the object to pick and its placement position, with each step rated independently, with the test performed solely by the GPT-4V module.
2) Execution success: Refers to the success rate of executing the entire task with the proposed system, with pick-and-place actions performed by the robot without human intervention.

GPT-4V generates pick-and-place instructions formatted as [object to pick, position to place] for task planning. For the planning success, each task is tested with 3 object arrangements on the tabletop workspace and 10 images per arrangement, totaling 30 images per step and up to 9 steps in total (S0-S8). For the execution success, each task undergoes 10 trials with different arrangements. The quantitative metric used for planning success is the success rate, calculated as the number of successes over the total number of trials. The quantitative metrics used for execution success are the correct task completion rate, *i.e.*, whether all objects are correctly placed [success, failure], and the correct placement rate, *i.e.*, the percentage of objects that are correctly placed.

Table I presents the mean planning success rate at each step for each task. T1 has the highest success rate, achieving a mean success rate equal to or higher than 0.93. For T2, the success rate increases once the first object has been placed in the box. In T3, the success rate is high from the beginning, always greater than 0.82. T4 has the lowest success rate, achieving a value of 0.70 after six objects have been placed. Indeed, the model misinterprets the task, recognizing an incorrect color-matching task instead of a sorting task.

Table II presents the system performance and execution success rates in correct completion and placement for each task. T1 shows that most errors are due to Detic misdetection. T2 achieves the best results for both correct task completion and correct placement rate. Most errors in T3 are due to GPT-4V, where the model fails to recognize the sorting task (*e.g.*, trying to find incorrect color-matching patterns).

TABLE III

STATISTICAL ANALYSIS OF COMPREHENSION EVALUATION VIA CONFIDENCE LEVEL COMPUTATION FOR EACH TASK (EXP. 2)

Task	Accuracy	Specificity	Precision
T1	1	-	1
T2	0.70	0.90	0.95
T3	0.78	0.67	0.94
T4	0.61	0.32	0.58

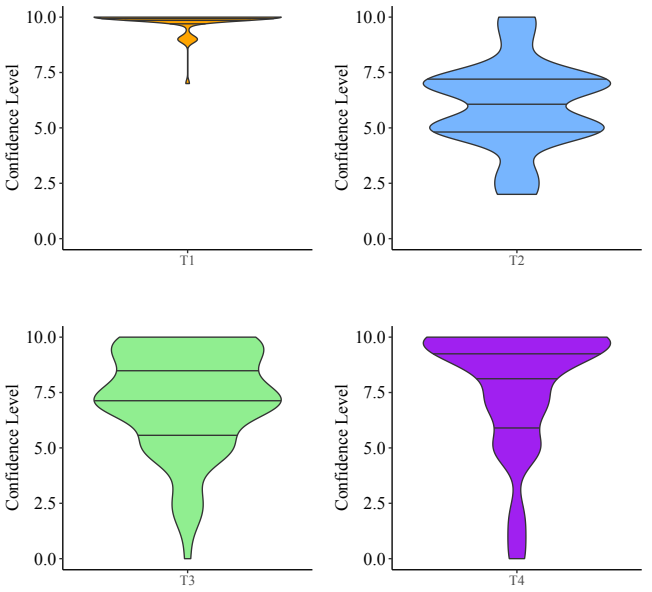


Fig. 5. Violin plots of the confidence levels for each task.

B. Comprehension Evaluation (Exp. 2)

This experiment aims to assess whether the proposed confidence level accurately reflects the performance of task planning, *i.e.*, whether GPT-4V's self-evaluation of task comprehension aligns with its actual planning success. For this purpose, each task is tested with 3 object arrangements and 4 images per arrangement, totaling 12 images per step. The evaluation criteria are accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$), specificity ($\frac{TN}{TN+FP}$), and precision ($\frac{TP}{TP+FN}$) across tasks, calculated using the confusion matrix that compares predictions to ground truth. A confidence level above the threshold is considered a positive prediction for successful task completion, while a level below the threshold is considered a negative prediction. We selected a threshold of 5 as it provides the best average performance. The recommended task planning acts as the ground truth, and is considered positive if the planning is correct and negative if it is not.

For T1, all the proposed plans are correct, and all the confidence levels are higher than 7, resulting in an accuracy and precision of 1. For T2 and T3, the accuracy and precision are similar, but the specificity is lower in T3. For T2, the distribution is bimodal, with peaks at confidence levels of 5 and 7. T3 shows a general trend where the number of occurrences increases with the confidence level, reaching a

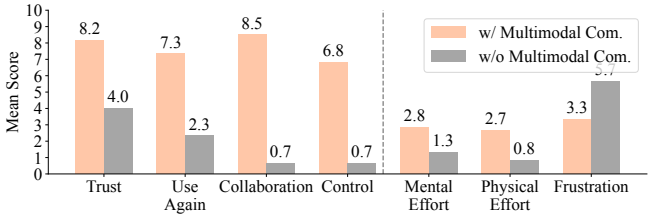


Fig. 6. Histogram comparing the qualitative results between the HRC experiment parts with and without multimodal communication. Metrics advantageous when high are shown on the left side of the dotted line, whereas those advantageous when low are displayed on the right side.

TABLE IV

p-VALUES OF THE QUALITATIVE METRICS FOR MULTIMODAL COMMUNICATION USING THE MANN–WHITNEY *U* TEST (EXP. 3)

Qual. Metric	Trust	Use Again	Collab.	Control	Mental Effort	Physical Effort	Frustr.
p-value	0.0001	0.0010	0.0001	0.0001	0.0589	0.0210	0.1700
Signif.	**	**	**	**		*	

Significant differences of $p < 0.05$ are denoted by $*$, $p < 0.01$ by $**$.

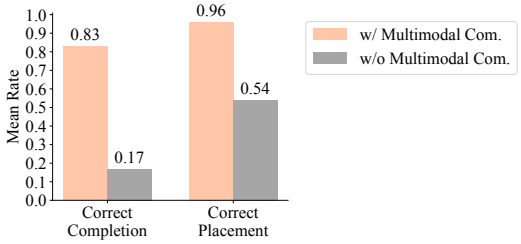


Fig. 7. Histogram comparing the quantitative results between the HRC experiment parts with and without multimodal communication.

TABLE V

p-VALUES OF THE QUANTITATIVE METRICS FOR TASK SUCCESS IN HRC USING THE MANN–WHITNEY *U* TEST (EXP. 3)

Quantitative Metric	Correct Completion	Correct Placement
p-value	0.03	0.02
Significance	*	*

Significant differences of $p < 0.05$ are denoted by $*$, $p < 0.01$ by $**$.

maximum at 7. T4 has both low precision and low specificity. For T4, the violin plot shows that the majority of confidence levels are high, indicating that the system is often confident about its task planning, but all metrics have the lowest values, meaning the model is incorrectly confident.

C. Multimodal Communication (Exp. 3)

This experiment aims to evaluate whether incorporating speech and gestures for communicating intention improves task success rates quantitatively and enhances human comfort qualitatively. Six human subjects were recruited to evaluate the collaboration in a two-part experiment: with and without multimodal communication. Each subject's role was to assist the robot in completing a specific task and to fill out a form to rate qualitative metrics on a scale from 0 to 10. We selected T3 for this experiment. The qualitative metrics assessed were feeling of trust, willingness to use again, sense of collaboration, feeling of control, mental effort, physical effort, and feeling of frustration. The quantitative metrics were again correct task completion and correct placement rates.

Fig. 6 shows the subjects' qualitative ratings for both experiment parts, with and without multimodal communication. Subjects rated HRC with multimodal communication higher across all metrics, except for mental and physical efforts, with trust receiving the highest scores overall. Table IV shows the *p*-values and significance of the results for each qualitative metric. Multimodal communication significantly improved HRC in terms of trust, willingness to use again, collaboration, and control. However, it was rated significantly lower in terms of physical effort. Fig. 7 and Table V present the results of the quantitative metrics. Multimodal communication during HRC significantly improved system performance in terms of both correct task completion and correct placement rate.

V. DISCUSSION

The proposed task planning system demonstrated that GPT-4V successfully interpreted T1, T2, and T3 but struggled with T4, which involved sorting with three colored mats. GPT-4V tended to incorrectly prioritize color matching, which negatively impacted both its performance and confidence in T4, misclassifying it as a color-matching task. This led to a failure in recognizing errors. However, introducing a confidence level was effective in detecting the most erroneous plans in T1, T2, and T3 by aligning GPT-4V's comprehension with its actual task planning.

Significant improvements in HRC were observed across four of the seven qualitative metrics: feeling of trust, willingness to use again, sense of collaboration, and feeling of control. The higher physical effort rating during HRC with multimodal communication, which included speech and gestures, is also reflected in the results. In terms of quantitative results, both the correct task completion and correct placement rates showed significant improvement when multimodal communication was used.

The results show that the proposed system could interpret implicit tasks in HRC and quantify its confidence in comprehension. However, clear limitations were also identified. GPT-4V's tendency to prioritize color matching posed a challenge in tasks where color was irrelevant, negatively impacting both task planning and confidence level assessments. Another potential limitation is the time required for the full execution of the system, which could be a constraint in more realistic HRC scenarios.

VI. CONCLUSION

We proposed a system designed to function reliably with non-predefined tasks implicitly instructed through HRC by focusing on three key aspects: 1) adaptive task planning using a multimodal LLM, 2) self-evaluation of the task comprehension, and 3) multimodal communication of intention to increase the HRC success rate and comfort. The

quantitative results showed that the proposed system can interpret implicitly instructed tabletop pick-and-place tasks through HRC, providing the next object to pick and the correct position to place it. The qualitative results showed that multimodal communication significantly enhances the feelings of trust and control, willingness to use again, and sense of collaboration during HRC. However, further evaluation of the proposed system should be performed in industrial or service environments to validate its effectiveness in more realistic task scenarios.

Although we selected GPT-4V for task planning due to its high performance, the rapid advancements in the multimodal LLM field suggest that other models may soon surpass it. For example, the recently introduced GPT-4o integrates text, audio, and vision into multimodal processing. This improvement could enable more complex confidence level computations, previously considered too time-consuming, and accelerate the proposed system execution. Nevertheless, spatial reasoning still remains challenging for multimodal LLMs. Ongoing research [33] aimed at improving spatial reasoning could significantly expand the utility of multimodal LLMs in task planning for HRC.

REFERENCES

[1] A. Brohan, *et al.*, "RT-1: Robotics Transformer for Real-World Control at Scale," Dec. 2022. Preprint: https://doi.org/10.48550/arXiv.2212.06817

[2] A. Mei, *et al.*, "GameVLM: A Decision-Making Framework for Robotic Task Planning based on Visual Language Models and Zero-Sum Games," May 2024. Preprint: https://doi.org/10.48550/arXiv.2405.13751

[3] A. Vaswani, *et al.*, "Attention Is All You Need," June 2017. Preprint: https://doi.org/10.48550/arXiv.1706.03762

[4] L. Ouyang, *et al.*, "Training Language Models to Follow Instructions with Human Feedback," Mar. 2022. Preprint: https://doi.org/10.48550/arXiv.2203.02155

[5] C. Y. Kim, *et al.*, "Understanding Large-Language Model (LLM)-powered Human-Robot Interaction," Mar. 2024. Preprint: https://doi.org/10.1145/3610977.3634966

[6] C. Zhang, *et al.*, "Large Language Models for Human–Robot Interaction: A Review," *Biomimetic Intelligence and Robotics*, vol. 3, no. 4, pp. 1–15 (100 131), Dec. 2023.

[7] M. Shridhar, *et al.*, "CLIPort: What and Where Pathways for Robotic Manipulation," Sept. 2021. Preprint: https://doi.org/10.48550/arXiv.2109.12098

[8] S. Hu, *et al.*, "LAMP: Leveraging Language Prompts for Multi-Person Pose Estimation," in *Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*, Detroit, United States, Oct. 2023, pp. 3759–3766.

[9] M. Ahn, *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," Apr. 2022. Preprint: https://doi.org/10.48550/arXiv.2204.01691

[10] W. Huang, *et al.*, "Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents," Jan. 2022. Preprint: https://doi.org/10.48550/arXiv.2201.07207

[11] B. Li, *et al.*, "Interactive Task Planning with Language Models," Oct. 2023. Preprint: https://doi.org/10.48550/arXiv.2310.10645

[12] N. Wake, *et al.*, "GPT-4V(ision) for Robotics: Multimodal Task Planning from Human Demonstration," Nov. 2023. Preprint: https://doi.org/10.48550/arXiv.2311.12015

[13] M. Kotelanski, *et al.*, "Methods to Estimate Large Language Model Confidence," Nov. 2023. Preprint: https://doi.org/10.48550/arXiv.2312.03733

[14] L. Li, *et al.*, "Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models," Feb. 2024. Preprint: https://doi.org/10.48550/arXiv.2402.12563

[15] L.-H. Lin, *et al.*, "Gesture-informed Robot Assistance via Foundation Models," Sept. 2023. Preprint: https://doi.org/10.48550/arXiv.2309.02721

[16] B. J. Grosz *et al.*, "Attention, Intentions, and the Structure of Discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, July 1986.

[17] S. Qu *et al.*, "Beyond Attention: The Role of Deictic Gesture in Intention Recognition in Multimodal Conversational Interfaces," in *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI 2008)*, Gran Canaria, Spain, Jan. 2008, pp. 237–246.

[18] N. Wake, *et al.*, "ChatGPT Empowered Long-Step Robot Control in Various Environments: A Case Application," *IEEE Access*, vol. 11, pp. 95 060–95 078, Aug. 2023.

[19] Y. Ding, *et al.*, "Robot Task Planning and Situation Handling in Open Worlds," Oct. 2022. Preprint: https://doi.org/10.48550/arXiv.2210.01287

[20] B. Purwanto, "Communication Science: The Role of Communication to Ensure Existence of Human," *Asian Journal of Management Sciences & Education*, vol. 7, no. 3, pp. 52–57, July 2018.

[21] S.-E. Fotinea, *et al.*, "The MOBOT Human-Robot Communication Model," in *Proceedings of 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2015)*, Gyor, Hungary, Oct. 2015, pp. 201–206.

[22] D. Yongda, *et al.*, "Research on Multimodal Human-Robot Interaction based on Speech and Gesture," *Computers & Electrical Engineering*, vol. 72, pp. 443–454, Nov. 2018.

[23] R. Stiefelhagen, *et al.*, "Natural Human-Robot Interaction using Speech, Head Pose and Gestures," in *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, vol. 3, Sendai, Japan, Sept. 2004, pp. 2422–2427.

[24] A. Oyama, *et al.*, "Exophora Resolution of Linguistic Instructions with a Demonstrative based on Real-World Multimodal Information," in *Proceedings of 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023)*, Busan, South Korea, Aug. 2023, pp. 2617–2623.

[25] R. Meena, *et al.*, "Integration of Gestures and Speech in Human-Robot Interaction," in *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, Slovakia, Dec. 2012, pp. 673–678.

[26] S. Paul, *et al.*, "Intent-based Multimodal Speech and Gesture Fusion for Human-Robot Communication in Assembly Situation," in *Proceedings of 21st IEEE International Conference on Machine Learning and Applications (ICMLA 2022)*, Nassau, Bahamas, Dec. 2022, pp. 760–763.

[27] S. Yoon, *et al.*, "Challenges in Deictic Gesture-Based Spatial Referencing for Human-Robot Interaction in Construction," in *Proceedings of the 38th IAARC International Symposium on Automation and Robotics in Construction (ISARC 2021)*, Dubai, United Arab Emirates, Nov. 2021, pp. 491–497.

[28] V. Schwind, *et al.*, "Up to the Finger Tip: The Effect of Avatars on Mid-Air Pointing Accuracy in Virtual Reality," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY 2018)*, Melbourne, Australia, Oct. 2018, pp. 477–488.

[29] X. Zhou, *et al.*, "Detecting Twenty-Thousand Classes Using Image-Level Supervision," in *Proceedings of 17th European Conference on Computer Vision (ECCV 2022)*, S. Avidan, *et al.*, Eds., Tel Aviv, Israel, Oct. 2022, pp. 350–368.

[30] A. Radford, *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022. Preprint: https://doi.org/10.48550/arXiv.2212.04356

[31] C. Lugaresi, *et al.*, "MediaPipe: A Framework for Perceiving and Processing Reality," in *Workshops of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, United States, June 2019, pp. 1–4.

[32] L. El Hafi, *et al.*, "Software Development Environment for Collaborative Research Workflow in Robotic System Integration," *RSJ Advanced Robotics (AR)*, vol. 36, no. 11, pp. 533–547, June 2022.

[33] B. Chen, *et al.*, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," Jan. 2024. Preprint: https://doi.org/10.48550/arXiv.2401.12168