Collaborative Conversation in Safe Multimodal Human-Robot Collaboration

Davide Ferrari, Andrea Pupa and Cristian Secchi

***Abstract*— In the context of Human-Robot Collaboration (HRC), it is crucial that the two actors are able to communicate with each other in a natural and efficient manner. The absence of a communication interface is often a cause of undesired slowdowns. On one hand, this is because unforeseen events may occur, leading to errors. On the other hand, due to the close contact between humans and robots, the speed must be reduced significantly to comply with safety standard ISO/TS 15066. In this paper, we propose a novel architecture that enables operators and robots to communicate efficiently, emulating human-to-human dialogue, while addressing safety concerns. This approach aims to establish a communication framework that not only facilitates collaboration but also reduces undesired speed reduction. Through the use of a predictive simulator, we can anticipate safety-related limitations, ensuring smoother workflows, minimizing risks, and optimizing efficiency. The overall architecture has been validated with a UR10e and compared with a state of the art technique. The results show a significant improvement in user experience, with a corresponding 23% reduction in execution times and a 50% decrease in robot downtime.**

I. Introduction

Human-Robot Collaboration (HRC) is rapidly gaining importance in various sectors of modern society, such as industry [1] [2], medicine [3] [4] and elder assistance [5] [6]. The success of HRC heavily relies on effective communication, akin to the significance of communication in Human-Human Collaboration (HHC) [7]. Ensuring seamless information sharing, accurate task execution in shared workspaces, and clear expression of intentions or requests between human and robot team members is paramount. In safety-focused scenarios, proactive communication becomes even more crucial, serving as one of the mechanisms to anticipate and avoid potential hazards or inefficiencies, fostering smooth collaboration while minimizing risks. While HRC shows promise in enhancing efficiency and safety across domains, existing approaches often struggle to achieve the dynamic, bidirectional, and proactive communication commonplace in human interactions, particularly in safety-related aspects. Communication in [8] and [9] enhances coexistence and trust, providing operators with insights into the intentions of robots. Meanwhile, [10] and [11] employ a multimodal communication approach, enabling unidirectional commands from the operator to the robot, enhancing safety by reducing misunderstandings. However, unidirectional communication lacks the ability for the robot to provide crucial feedback, as seen in [12], which introduces bidirectional communication where the robot can request safety constraint relaxation using Control Barrier Functions (CBFs). This approach ensures that the robot can express its needs to the operator, enhancing both safety and task efficiency. Despite these efforts, existing approaches often prioritize communication or safety separately, leading to a critical gap where safety is not explicitly integrated into communication strategies. Approaches like [13]–[16] prioritize safety but lack integration with dedicated communication channels, turning safety into a barrier that may cause the robot to stop or slow down without explicitly guaranteeing compliance with ISO/TS 15066 safety standards [17]. This gap underscores the need for further research and development to seamlessly integrate safety and communication, meeting regulatory requirements and ensuring safe collaboration in diverse applications. In this article, we introduce an innovative HRC approach that prioritizes natural conversation while utilizing a predictive simulator to proactively anticipate potential safety issues, all while adhering to ISO safety standards. Our architectural framework views human and robot team members as equals, enabling an intuitive bidirectional communication to facilitate dynamic information exchange, closely resembling human-to-human conversations. Through bidirectional communication, the system engages with the user to find solutions, minimizing risks while maintaining high efficiency and ensures smoother workflows avoiding disruptions and preemptively addressing safety concerns.

Thus, the contributions of this paper are:

- An innovative architecture for Bidirectional Multimodal Communication that enables natural dialogues between humans and robots.
- A predictive simulator and the seamless integration of safety considerations into the conversation, proactively addressing potential slowdowns or blockages to enhance collaboration efficiency.
- An experimental validation by comparing our proposed architecture to the state-of-the-art, where safety functions as a low-level layer and is not an integral part of communication.

This paper is structured as follows: in Section II, we address the Problem Statement. Section III presents the Proposed Architecture, while Section IV delves into the Safety Layer and the Predictive Simulator. Experimental Validation, Implementation Details, and Analysis of the Results are discussed in Section V, and in Section VI, we draw our Conclusions.

D. Ferrari, A. Pupa and C. Secchi are with the Department of Sciences and Methods of Engineering, University of Modena and Reggio Emilia, Italy {davide.ferrari95, andrea.pupa, cristian.secchi}@unimore.it

II. Problem Statement

Consider a Human-Robot Collaboration scenario, where a 6-degree-of-freedom (6-DOF) velocity-controlled manipulator, modeled as:

$$\dot{\mathbf{q}} = \mathbf{u} \tag{1}$$

where $\dot{\mathbf{q}} \in \mathbb{R}^n$ denotes the joints velocities and $\mathbf{u} \in \mathbb{R}^n$ represents the controller input, is required to cooperate and establish communication with a human operator to achieve a common objective. In this context, humans and robots collaborate to execute a cooperative assembly task within a shared workspace. The robot's primary role is to assist the operator to complete the task efficiently. Utilizing a set of sensors, real-time monitoring of both the human operator's position and the objects in the workspace becomes possible. This real-time monitoring forms the basis for planning safe trajectories $\mathbf{q}_{des}(t) \in \mathbb{R}^n$, which originate from an initial configuration $\mathbf{q}_{des}(t_i) = \mathbf{q}_i \in \mathbb{R}^n$ and extend to a desired final configuration $\mathbf{q}_{des}(t_f) = \mathbf{q}_f \in \mathbb{R}^n$. These trajectories must adhere to the ISO/TS 15066 regulations [17], which imposes constraints on the maximum speed in the direction of the operator [18]:

$$\begin{aligned} v_{rh}(t) \leq & \sqrt{v_h(t)^2 + (a_{max}T_r)^2 - 2a_{max}K(t)} + \\ & - a_{max}T_r - v_h(t), \end{aligned} \tag{2}$$

where $K(t) = C + Z_d + Z_r - S_p(t)$. $v_{rh}(t) \in \mathbb{R}$ and $v_h(t) \in \mathbb{R}$ are the scalar velocity of the robot towards the human operator and the scalar velocity of the human operator towards the robot, respectively. $a_{max} \in \mathbb{R}$ is the maximum deceleration and $T_r \in \mathbb{R}$ is the robot reaction time. C is the intrusion distance, i.e. the distance that a part of the body can intrude into the sensing field before it is detected, while Z_d and Z_r are the position uncertainties of the human operator inside the workspace and of the robot system, respectively. Lastly, S_p represents the protective separation distance.

To ensure compliance with safety standards while maintaining the overall path integrity, a strategy is employed that explicitly isolates the velocity magnitude along the trajectory. This is achieved through a path-velocity decomposition method, involving the manipulation of the derivative $\dot{s}$ of the curvilinear abscissa s that parameterizes the geometric path $\mathbf{q}_{des}(s(t))$ as follows:

$$\mathbf{q}_{des}(t) = \mathbf{q}_{des}(s(t)) \qquad t \in \left[t_i, t_f\right], \tag{3}$$

$$\dot{\mathbf{q}}_{des}(t) = \mathbf{q}'_{des}(s(t))\dot{s} \qquad t \in \left[t_i, t_f\right], \tag{4}$$

However, since all robot tasks must be implemented safely, this often translates into inefficient behavior, causing undesired slowdowns or blockages. In a HHC scenario, the two actors would anticipate potential issues by discussing and exchanging valuable information to avoid them. To emulate this, it is first necessary to implement a multimodal conversation architecture capable of handling multiple communication channels and enabling the exchange of information in a simple, natural, and dynamic manner. Secondly, the robot needs to be equipped with an algorithm that allows it to assess and predict the possible emergence of such issues, enabling it to converse with the operator and find a solution.

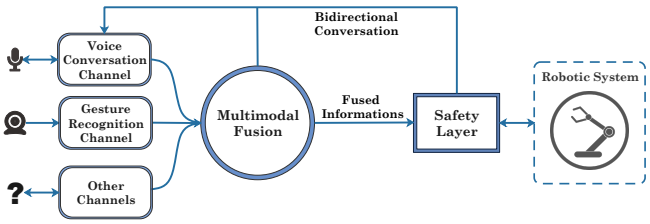

Fig. 1. Proposed Architecture

This work aims to address these issues by enabling multimodal conversations, which, through effective communication and predictive simulation, aim to prevent the emergence of potential future problems. To achieve this, we propose a safe-integrated architecture that:

- enables multimodal conversations between humans and robots to ensure efficient and natural information exchange.
- anticipates the emergence of potential issues or slowdowns.
- ensures compliance with regulations during the execution of the robot's trajectories.

III. Proposed Architecture

The proposed architecture, depicted in Fig. 1, integrates multiple communication channels, both unidirectional (e.g., gesture recognition) and bidirectional (e.g., voice conversation), contributing inputs to a multimodal fusion core. This core receives and synchronizes information from diverse channels, generating hybrid data to reconstruct the received multimodal communication. Additionally, a crucial safety layer manages robot inputs, simulates potential issues through a simulation algorithm, communicates errors, obstacles, or blockages, and proactively initiates conversations to address potential hazards. Our Human-Robot Collaboration (HRC) approach focuses on facilitating natural multimodal conversations, resembling human-to-human dialogue fluidity, and preventing the activation of constraints imposed by ISO safety standards through communication. In our architectural framework, communication channels can be interchangeably used by both humans and robots to initiate conversations, exchange information, coordinate tasks, and make requests, fostering a dynamic interaction model. To enable bidirectional communication, we implemented a core multimodal fusion system that seamlessly integrates various communication channels, such as voice and gestures, creating an efficient and versatile communication environment. Safety standards are seamlessly integrated within this communicative framework to proactively address potential safety issues. A predictive simulator anticipates potential slowdowns or blockages in the robot's trajectory, engaging in proactive dialogue with the operator to preemptively address safety concerns, prevent disruptions, and ensure smoother workflows while minimizing risks. Our architecture uniquely combines effective communication and safety standards as integral components, distinguishing it from existing approaches [12]. Unlike traditional methodologies that often treat safety as a lower-level

layer, we elevate it to an equal standing with communication, ensuring compliance with ISO/TS 15066 safety standards. This approach establishes safety as a core component of our communication strategies with operators.

A. Multimodal Fusion

Our multimodal fusion approach is grounded in the architecture outlined in [19]. It utilizes a combination of diverse communication channels seamlessly merged through a multimodal fusion algorithm. The primary objective of this integration is to form a unified and comprehensive representation of the communicated message, leveraging inputs from various communication channels. The process of multimodal fusion involves two key steps. Initially, information originating from these communication channels is collected and managed by a time manager inspired by the recognition lines discussed in [20]. This crucial component is responsible for synchronizing and consolidating data from different sources into a singular tensor. Following this, a neural network classifier processes the tensor and generates the fused multimodal command, which is essentially a coherent and holistic representation of the received communication. The decision to employ a neural network as the classifier is driven by its intrinsic ability to learn and adapt to the complex, non-linear nature of multimodal data. Compared to other types of classifiers, neural networks often demonstrate greater flexibility in modeling intricate relationships and capturing complex patterns in data. This adaptability is particularly advantageous when handling signals from diverse communication channels. Furthermore, their adaptability allows seamless integration of new inputs and channels through the incorporation of additional training data. This flexibility ensures the system's responsiveness to evolving communication scenarios, making it a versatile and effective choice for real-world applications. As outlined in Algorithm 1, upon the arrival of new information from either the voice or gesture channel, a temporal window $\mathcal{W}$ is initiated. Within this window, any additional information received by each communication channel, denoted as $(X_1, X_2, \ldots, X_n)$, is concatenated into a tensor $\mathcal{T} = [X_1, X_2, \ldots, X_n]$ of predefined length, applying the zero-padding technique to fill the available slots for each communication channel and ensuring conformity with the input dimension of the neural network. Subsequently, the synchronized tensor $\mathcal{T}$ is given as input to a pre-trained neural classifier $\mathcal{N}$, represented by a function $\mathcal{M} = F(\mathcal{T})$, which generates a fused multimodal command $\mathcal{M}$.

This neural network is structured as a sequential model consisting of three linear layers interspersed with Rectified Linear Unit (ReLU) [21] activation function $\sigma(x) = \max(0, x)$ that introduces non-linearity into the model, allowing it to learn complex patterns in the data. The network is parameterized by a set of weights W and biases b, which are fine-tuned during the training process. The classification function can be expressed as:

$$\mathcal{M} = \sigma(W \cdot \mathcal{T} + b) \tag{5}$$

The input and output sizes are dependent on the nature of the communication channels and the experimental setup. In our case, the input size is set to 4, reflecting the constraint of a narrow temporal window where it is unlikely to receive more than one vocal and gestural command and the output size is configured to 15 classes, each corresponding to a specific meaning in the context of the experiment, such as *"place object there"*, *"re-plan trajectory*, etc. Additionally, each hidden layer has a hidden size of 64 neurons, contributing to the network's capacity to capture complex relationships within the data.

Algorithm 1 Multimodal Fusion Algorithm

Require: Vectors G (*Gesture*) and V (*Voice*) information
Ensure: Multimodal Command $\mathcal{M}$
$\mathcal{T} \leftarrow$ empty tensor
Recognition Time: $R_T \leftarrow 2s$
if new value of G or V is received **then**
  Open a *Temporal Window* $\mathcal{W}$
  $\mathcal{T} \leftarrow$ Received Value (G or V)
  while $\mathcal{W} < R_T$ **do**
    if new gesture/voice G or V is received **then**
      Fill tensor $\mathcal{T}$ with new G or V
    end if
  end while
  Pass $\mathcal{T}$ in Classifier $\mathcal{N} \rightarrow$ Multimodal Command $\mathcal{M}$
  Send $\mathcal{M}$ to the Safety Layer
end if

We constructed a custom dataset collected experimentally by classifying various combinations of inputs into classes based on their meanings in the context of the experiment. Each data point in the dataset is represented as $(\mathcal{T}_i, Y_i)$, where $\mathcal{T}_i$ corresponds to the synchronized tensor from multiple communication channels, and Y_i is the corresponding ground truth characterizing the desired fused representation. The training process was conducted using the Stochastic Gradient Descent (SGD) optimization algorithm with a batch size of 64 and a learning rate of 4e-3 to minimize a loss function L that quantifies the difference between the predicted $\mathcal{M}$ and the ground truth Y:

$$L = \sum_i \mathcal{L}(\mathcal{M}_i, Y_i), \tag{6}$$

where $\mathcal{L}$ is a suitable loss function, such as cross-entropy. Additionally, we implemented the early-stopping technique with a patience of 100 epochs to monitor the trend of the loss function and stop the training prematurely to prevent overfitting.

B. Safety

Safety is maintained through a dedicated layer, outlined in Sec. IV, which is responsible for both protecting the human operator and avoid undesired slowdowns. The ISO/TS 15066 sets a maximum limit on the relative human-robot speed, as indicated in equation (2). However, adhering to this constraint might result in a too conservative robot behaviour. Thanks to the bidirectional communication channel, the robot

can communicate its intentions to the human operator, preventing undesired speed reduction and enhancing the overall collaboration performance.

IV. Safety Layer

The safety layer has two main goals. Firstly, it ensures that the robot behavior complies with safety standards. Secondly, it anticipates and communicates future speed reductions to improve performance. Thanks to the proposed approach, the robot is capable of warning the human operator about its intention, emulating real human-human communication. Indeed, when two human operators collaborate, it is common for one of the two to ask the other to move since it has to reach the same area of the workspace. This translates into an improvement of the mutual communication and comprehension of the two different agents, without compromising safety. To achieve this, the safety layer computes for each task a collision-free trajectory $\mathbf{q}_{des}(t)$, which is decomposed through a path-velocity decomposition as detailed in (3)-(4). Subsequently, the trajectory is forwarded to two different components, each of which is responsible for one of the goals: online safety and predictive simulator.

A. Online Safety

The online safety of the human operator is guaranteed by a velocity scaling algorithm which was initially proposed in [22]. In particular, starting from the trajectory, the safety layer solves online the following optimization problem:

$$\begin{aligned} \min_{\alpha} \quad & -\alpha, \\ \text{s.t.} \quad & \\ & J_{r_i}(\mathbf{q})\mathbf{q}'(s)\dot{s}\alpha \leq v_{max_i} \quad \forall i \in \{1,\dots,n\}, \\ & \dot{\mathbf{q}}_{min} \leq \mathbf{q}'(s)\dot{s}\alpha \leq \dot{\mathbf{q}}_{max}, \\ & \ddot{\mathbf{q}}_{min} \leq \frac{\mathbf{q}'(s)\dot{s}\alpha - \dot{\mathbf{q}}}{T_r} \leq \ddot{\mathbf{q}}_{max}, \\ & 0 \leq \alpha \leq 1. \end{aligned} \tag{7}$$

$\alpha \in \big[0,1\big]$ is the optimization variable and represents the scaling factor. $J_{r_i}(\mathbf{q}) \in \mathbb{R}^{1\times n}$ is a *modified jacobian* that takes into account only the scalar velocity of the i-th link towards the human operator, see [22]. v_{max_i} is the velocity limit imposed by the ISO/TS 15066 [17]. $\dot{\mathbf{q}}_{min} \in \mathbb{R}^n$ and $\dot{\mathbf{q}}_{max} \in \mathbb{R}^n$ are the joint velocity lower bounds and the joint velocity upper bounds, respectively. While $\ddot{\mathbf{q}}_{min} \in \mathbb{R}^n$ and $\ddot{\mathbf{q}}_{max} \in \mathbb{R}^n$ are the acceleration limits. $\dot{\mathbf{q}} \in \mathbb{R}^n$ is the actual robot velocity and T_r is the robot execution time.

B. Predictive Simulator

The predictive simulator has the goal of predicting and avoiding undesired modulation of the speed. Indeed, according to (7), when the human operator and the robot are close and the robot is going towards the human operator, the robot speed is scaled over the path to ensure compliance with the safety standards. However, these situations may cause a stop of the robot for a huge time, i.e. $\alpha = 0$ until the human operator moves away. Thus, it would be more beneficial to communicate that the robot will decrease its speed so that the human operator can decide to move away and avoid useless stuck.

The predictive simulator strategy is implemented according to Algorithm 2. The algorithm requires as input the parametrized trajectory and its derivative, namely $\mathbf{q}_{des}(s(t))$ and $\mathbf{q}'_{des}(s(t))$. It immediately sets to $false$ the variable $triggered$, which indicates if the warning message has been sent, and finds what is the last robot configuration along the path, i.e. $\mathbf{q}_{end}$ (Lines 1-2). Subsequently, it starts a while loop in which it initially reads the real robot position along the planned path and initializes the state of the virtual robot $\mathbf{q}_{virt}$. Then, the algorithm initializes the variable End_{traj}, which is used to check if the virtual robot has concluded the planned path, the virtual time T_{virt}, and it computes the remaining trajectory time T_{rem}, i.e. the ideal time that the robot needs to conclude the trajectory. Then, the algorithm starts an inner while loop. In this loop, the predictive simulator continuously updates the information regarding the human operator, i.e. position, velocity and human-robot distance. This information can be computed by exploiting standards techniques already available in the literature, e.g. distance between capsules [23] and human motion prediction [24]. Subsequently, the algorithm computes the optimal scaling factor of the virtual robot α_{virt} by solving the problem in (7) and integrates the dynamics of the virtual robot (Line 12). Such integration can be achieved by exploiting standard methods, e.g. forward Euler [25]. Then, the algorithm increments T_{virt} and checks if the virtual robot has reached the end of the path. If this is the case, the predictive simulator exits from the inner while and restarts the check from the actual robot configuration (Line 15). If the path is not concluded, the algorithm checks if T_{virt} significantly

Algorithm 2 Predictive Simulator

Require: $\mathbf{q}_{des}(s(t)), \mathbf{q}'_{des}(s(t))$
1: $triggered \leftarrow false$
2: $\mathbf{q}_{end} \leftarrow getEnd(\mathbf{q}_{des}(s(t)))$
3: **while** not $triggered$ **do**
4: $\quad \mathbf{q}_{virt} \leftarrow getRealState(\mathbf{q}_{des}(s(t)))$
5: $\quad End_{traj} \leftarrow false$
6: $\quad T_{virt} \leftarrow 0$
7: $\quad T_{rem} \leftarrow getDuration(\mathbf{q}_{des}(s(t)), \mathbf{q}_{virt})$
8: $\quad$ **while** not End_{traj} **do**
9: $\quad\quad \dot{s} \leftarrow getSpeed(\mathbf{q}'_{des}(s(t)), \mathbf{q}_{virt})$
10: $\quad\quad H_{info} \leftarrow getHumanData()$
11: $\quad\quad \alpha_{virt} \leftarrow solveOpt(\dot{s}, \mathbf{q}_{virt}, H_{info})$
12: $\quad\quad \mathbf{q}_{virt} \leftarrow integrate(\dot{s}, \mathbf{q}_{virt}, \alpha_{virt})$
13: $\quad\quad T_{virt} \leftarrow T_{virt} + T_r$
14: $\quad\quad$ **if** $\mathbf{q}_{virt} = \mathbf{q}_{end}$ **then**
15: $\quad\quad\quad End_{traj} \leftarrow true$
16: $\quad\quad$ **end if**
17: $\quad\quad$ **if** $checkTime(T_{virt}, T_{rem})$ **then**
18: $\quad\quad\quad sendSignal()$
19: $\quad\quad\quad triggered \leftarrow true$
20: $\quad\quad\quad$ **break**
21: $\quad\quad$ **end if**
22: $\quad$ **end while**
23: **end while**

exceeds the remaining time (Line 17). If this is the case, it means that the robot will have to reduce the speed too much, with a consequent reduction in the performance. Therefore, the algorithm communicates the warning message and stops (Lines 18-19). The algorithm is activated every time the safety layer computes a new trajectory, this is because if the human operator decides to not move it is not necessary to communicate again possible speed reduction.

It is worth underline that, to achieve its functionality,the predictive simulator needs to operate at a frequency much higher than the robot controller. This is doable because the robot is modeled as a kinematic system, as detailed in Sec. II.

V. Experimental Validation

The experimental validation[1] of our bidirectional multimodal communication control architecture involved a comparative study between the proposed architecture and a state-of-the-art bidirectional communication system by Ferrari et al. [12], which simply adheres to ISO safety standards through a low-level layer. The study involved 12 participants, aged between 20 and 30 years, with an equal gender distribution and varying levels of experience with robotic systems, ranging from first-time encounters to several years of experience with collaborative robots, in order to ensure a representative sample. To mitigate potential learning biases, each participant performed both versions of the experiment, and the execution order was randomly determined for each individual, minimizing the impact of experiment repetition on the results. In this experimental scenario (Fig. 2), participants engaged in a collaborative assembly task where they assembled a set of LEGO components. The robot, a UR10e collaborative manipulator, played a supportive role by providing the necessary components to the operator. The experiments encompassed a variety of interactions, including requests for assistance from the robot, error notifications, and anticipation of potential issues. Central to our experiment was the communication between the robot and the operator, facilitated through both vocal conversation and a 3D gesture recognition channel. The robot can initiate conversations with the operator to communicate events and work together to find or propose solutions to emerging challenges. Conversely, the operator can send input to the robot through both the vocal conversation and gesture recognition channels. The multimodal fusion algorithm fused information from the multiple communication channels, enabling the creation of commands like *"place the object in that area"* with the operator indicating the desired area through a *"Point-At"* gesture, extrapolating the target zone by interpolating the coordinates of the elbow, wrist, and finger using a skeletonization algorithm. The primary objective of these experiments was to evaluate the effectiveness of multimodal conversation compared to common bidirectional communication while ensuring compliance with safety regulations.

[1]In the accompanying video, a sneak peek of some key parts of the experiment is showcased.

Fig. 2. Setup of the Experiment

Additionally, we assessed its impact on collaboration efficiency and safety, including a predictive simulator to predict and preemptively address potentially hazardous situations through communication with the operator. Our goal was to highlight the advantages and differences brought about by our innovative approach.

The following sections present a detailed discussion of the architecture implementation and the experimental results, providing insights into the benefits and enhancements achieved by our multimodal conversation architecture.

A. Implementation Details

The architecture was developed using the *ROS* framework [26], with components organized into independent nodes to ensure modularity. For the voice communication channel, we created a custom *Amazon Alexa skill*, designed using *Alexa Conversations* [27], a Deep Learning-based approach that employs API calls to manage multi-turn dialogues between Alexa and the user, resulting in more natural and human-like interactions. The skill's back-end was locally developed in Python (non-Alexa-hosted), enabling seamless integration with ROS by utilizing *Microsoft Azure's HTTP Trigger Functions*. We employed *ngrok* to expose the back-end and establish a connection to the front-end via HTTPS tunneling. To enable bidirectional communication and allow the robot to initiate conversations, we integrated *Node-RED* [28], a web service for logical path programming. *Node-RED* provides direct interaction with Alexa APIs, empowering the robot to report errors, manage events, and start conversations by invoking specific dialogue APIs, thereby facilitating the exchange of information and task execution. The gesture recognition channel was established using the *Holistic* landmarks detection solution API from *MediaPipe* [29], a framework developed by Google that offers a comprehensive suite of libraries and tools for the application of artificial intelligence (AI) and machine learning (ML) techniques. Holistic combines elements of pose, face, and hand landmarks to create a unified landmarking system for the human body, operating in real-time on a continuous stream of images. The landmarks extracted from each image are encoded into a tensor to represent 3D gestures, which are then fed into a neural network classifier. This classifier consists of an LSTM (Long Short-term Memory) layer, followed by several fully connected layers, designed to classify gestures based on the extracted landmarks. The neural network model was

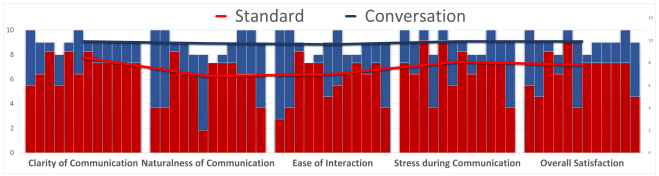

Fig. 3. Questionnaire Results

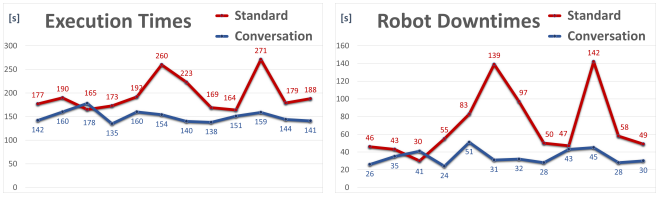

Fig. 4. Execution Times and Robot Downtime

trained using a dataset of communicative gestures specifically curated for human-robot collaboration [30].

B. Analysis of the Results

To assess the effectiveness of the architecture, we measured the execution times and downtime of the robot during both versions of the experiment. Given that compliance with ISO safety standards is ensured in both experiments, we aim to use these metrics to valuate how well our architecture can predict and prevent slowdowns or safety stops through communication, specifically those arising from excessive proximity to the operator. Additionally, for each experiment, we administered a questionnaire to the participants, consisting in five ratings on a scale from 0 to 10, covering *Clarity of Communication*, *Naturalness of Communication*, *Ease of Interaction*, *Stress during Communication*, and *Overall Satisfaction*. The graph in Fig. 3 displays the questionnaire results, with values from the comparative experiment shown in red and those from the architecture proposed in this article shown in blue. The results indicate a significant difference: the proposed architecture has an average score of approximately 9/10, while Ferrari at al. [12] approach averages around 7/10. This suggests a notable improvement in the user experience when using the proposed architecture, both in terms of clarity and ease of use, as well as in terms of reduced stress due to more natural communication.

The results of the experiment *Execution Times* and *Robot downtime* are depicted in Fig. 4. It is evident that the utilization of the proposed architecture leads to a significant enhancement in job performance, with an average completion time of 150 seconds compared to the standard communication's average of 196 seconds. Furthermore, robot downtime also dramatically decreases with the use of the proposed architecture, reducing from an average of 70 seconds to 35 seconds. Thanks to the integrated predictive simulator and communication, potential blockages and slowdowns were anticipated and avoided, leading to faster execution of the required tasks and reduced robot downtime.

TABLE I
ANOVA SUMMARY AND RESULTS - EXECUTION TIMES

Groups	Count	Sum	Average	Variance
Standard Communication	12	2351	195.92	1314.45
Multimodal Conversation	12	1802	150.17	157.42

F	P-value	F crit.
17.065	0.00044	4.30095

TABLE II
ANOVA SUMMARY AND RESULTS - ROBOT DOWNTIME

Groups	Count	Sum	Average	Variance
Standard Communication	12	839	69.92	1407.90
Multimodal Conversation	12	414	34.5	73

F	P-value	F crit.
10.164	0.00425	4.30095

TABLE III
ANOVA SUMMARY AND RESULTS - QUESTIONNAIRE RESULTS

Groups	Count	Sum	Average	Variance
Standard Communication	5	35	7	0.4167
Multimodal Conversation	5	45	9	0.0139

F	P-value	F crit.
46.452	0.00014	5.31765

To evaluate the statistical significance of the experiment and validate our findings, we conducted a single-factor ANOVA test for execution times, robot downtime and questionnaire results. Tables I, II and III summarize the collected data, including mean values, sums, and variances, categorized by the two experiments. The results demonstrate that in all the ANOVA tests, the calculated F-value is significantly higher than the F-critical value, and the p-value is well below the significance level (alpha = 0.05), confirming a statistically significant difference between the two experiments.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a novel bidirectional multimodal communication architecture designed to enhance human-robot collaboration in shared workspaces while prioritizing safety. Our architecture enables robots and human operators to engage in natural and effective conversations, mirroring the fluidity of human-to-human dialogue while adhering to ISO safety standards. The experimental validation demonstrates the effectiveness of the architecture, achieving great collaboration efficiency and a more user-friendly and natural interaction experience.

Future works will aim at expand the communication capabilities of our architecture by introducing new communication channels and modalities, allowing for even richer and more versatile interactions between humans and robots. Subsequently, we plan to incorporate vision AI algorithms to enhance error handling and object detection. This addition

will empower robots to better perceive and react to their environment. Furthermore, it is possible to implement a real-time user monitoring system capable of assessing the operator's status, stress levels, and concentration. This monitoring will enable the robot to adapt its behavior to better assist the operator and proactively address any emerging issues.

REFERENCES

[1] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, no. 4, 2019. [Online]. Available: https://www.mdpi.com/2218-6581/8/4/100

[2] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957415818300321

[3] T. Haidegger, S. Speidel, D. Stoyanov, and R. M. Satava, "Robot-assisted minimally invasive surgery—surgical robotics in the data age," *Proceedings of the IEEE*, vol. 110, no. 7, pp. 835–846, 2022.

[4] Y. Shen, D. Guo, F. Long, L. A. Mateos, H. Ding, Z. Xiu, R. B. Hellman, A. King, S. Chen, C. Zhang, and H. Tan, "Robots under covid-19 pandemic: A comprehensive survey," *IEEE Access*, vol. 9, pp. 1590–1615, 2021.

[5] J. Mišeikis, P. Caroni, P. Duchamp, A. Gasser, R. Marko, N. Mišeikienė, F. Zwilling, C. de Castelbajac, L. Eicher, M. Früh, and H. Früh, "Lio-a personal robot assistant for human-robot interaction and care applications," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5339–5346, 2020.

[6] M. R. Lima, M. Wairagkar, M. Gupta, F. Rodriguez y Baena, P. Barnaghi, D. J. Sharp, and R. Vaidyanathan, "Conversational affective social robots for ageing and dementia support," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1378–1397, 2022.

[7] K. Mojtahedi, B. Whitsell, P. Artemiadis, and M. Santello, "Communication and inference of intended movement direction during human–human physical interaction," *Frontiers in Neurorobotics*, vol. 11, 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2017.00021

[8] S. Grushko, A. Vysocký, P. Oščádal, M. Vocetka, P. Novák, and Z. Bobovský, "Improved mutual understanding for human-robot collaboration: Combining human-aware motion planning with haptic feedback devices for communicating planned trajectory," *Sensors*, vol. 21, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3673

[9] R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal, "Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction," *Robotics and Computer-Integrated Manufacturing*, vol. 61, p. 101830, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584518303351

[10] E. Rosen, D. Whitney, M. Fishman, D. Ullman, and S. Tellex, "Mixed reality as a bidirectional communication interface for human-robot interaction," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 431–11 438.

[11] S. Liu, L. Wang, and X. V. Wang, "Symbiotic human-robot collaboration: multimodal control using function blocks," *Procedia CIRP*, vol. 93, pp. 1188–1193, 2020, 53rd CIRP Conference on Manufacturing Systems 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212827120305874

[12] D. Ferrari, F. Benzi, and C. Secchi, "Bidirectional communication control for human-robot collaboration," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2022, p. 7430–7436. [Online]. Available: https://doi.org/10.1109/ICRA46639.2022.9811665

[13] F. Benzi, F. Ferraguti, and C. Secchi, "Energy tank-based control framework for satisfying the iso/ts 15066 constraint," *arXiv preprint arXiv:2304.14059*, 2023.

[14] F. Ferraguti, M. Bertuletti, C. T. Landi, M. Bonfè, C. Fantuzzi, and C. Secchi, "A control barrier function approach for maximizing performance while fulfilling to iso/ts 15066 regulations," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5921–5928, 2020.

[15] A. Palleschi, M. Hamad, S. Abdolshah, M. Garabini, S. Haddadin, and L. Pallottino, "Fast and safe trajectory planning: Solving the cobot performance/safety trade-off in human-robot shared environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5445–5452, 2021.

[16] N. Lucci, B. Lacevic, A. M. Zanchettin, and P. Rocco, "Combining speed and separation monitoring with power and force limiting for safe collaborative robotics applications," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6121–6128, 2020.

[17] I. O. for Standardization, "Iso/ts 15066:2016(e). robots and robotic devices–collaborative robots," International Organization for Standardization, Geneva, CH, Technical Specification, Feb. 2016.

[18] A. Pupa, M. Arrfou, G. Andreoni, and C. Secchi, "A human-centered dynamic task scheduling and safe task execution approach for human-robot collaboration scenarios," in *Human-Robot Collaboration: Unlocking the potential for industrial applications*, ser. Control, Robotics and Sensors. Institution of Engineering and Technology, 2023, pp. 105–130. [Online]. Available: https://digital-library.theiet.org/content/books/10.1049/pbce134e_ch6

[19] D. Ferrari, A. Pupa, A. Signoretti, and C. Secchi, "Safe multimodal communication in human-robot collaboration (forthcoming)," in *International Workshop on Human-Friendly Robotics*. Springer, 2023.

[20] J. Cacace, A. Finzi, and V. Lippiello, "A robust multimodal fusion framework for command interpretation in human-robot cooperation," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 372–377.

[21] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[22] A. Pupa, M. Arrfou, G. Andreoni, and C. Secchi, "A safety-aware kinodynamic architecture for human-robot collaboration," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4465–4471, 2021.

[23] F. Ferraguti, C. T. Landi, S. Costi, M. Bonfè, S. Farsoni, C. Secchi, and C. Fantuzzi, "Safety barrier functions and multi-camera tracking for human–robot shared environment," *Robotics and Autonomous Systems*, vol. 124, p. 103388, 2020.

[24] H. Liu and L. Wang, "Human motion prediction for human-robot collaboration," *Journal of Manufacturing Systems*, vol. 44, pp. 287–294, 2017.

[25] B. Biswas, S. Chatterjee, S. Mukherjee, and S. Pal, "A discussion on euler method: A review," *Electronic Journal of Mathematical Analysis and Applications*, vol. 1, no. 2, pp. 2090–2792, 2013.

[26] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[27] A. Acharya, S. Adhikari, S. Agarwal, V. Auvray, N. Belgamwar, A. Biswas, S. Chandra, T. Chung, M. Fazel-Zarandi, R. Gabriel, S. Gao, R. Goel, D. Hakkani-Tur, J. Jezabek, A. Jha, J.-Y. Kao, P. Krishnan, P. Ku, A. Goyal, C.-W. Lin, Q. Liu, A. Mandal, A. Metallinou, V. Naik, Y. Pan, S. Paul, V. Perera, A. Sethi, M. Shen, N. Strom, and E. Wang, "Alexa conversations: An extensible data-driven approach for building task-oriented dialogue systems," 2021.

[28] C. OpenJS Foundation, "Node-red." [Online]. Available: https://nodered.org

[29] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[30] J. Tan, W. P. Chan, N. L. Robinson, E. A. Croft, and D. Kulić, "A proposed set of communicative gestures for human robot interaction and an rgb image-based gesture recognizer implemented in ros," *ArXiv*, vol. abs/2109.09908, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237581181