Ref. Ares(2024)8762912 - 09/12/2024

This project is funded by the European Union
G.A No 101016906

Project Number: 101016906
Start Date of Project: 2021/01/01
Duration: 48 months

Type of document D9.2 – V1.0

Validation of Human-Robot Interaction

Dissemination level	PUBLIC
Submission Date	2024-12-08
Work Package	WP9
Task	T9.2, T9.3, T9.4
Type	Report
Version	1.0
Author	UNIROMA1, UNICLAM, UPC
Approved by	UNIROMA3

DISCLAIMER:
The sole responsibility for the content of this deliverable rests with the authors. It does not necessarily reflect the opinion of the European Union. Neither the Directorate-General for Communications Networks, Content and Technology, Artificial Intelligence and Digital Industry nor the European Commission are responsible for any use that may be made of the information

This project is funded by the European Union
G.A No 101016906

Executive Summary

This deliverable describes the validation of the human-robot interaction functionalities during harvesting and pruning as per Tasks 9.2, 9.3 and 9.4. Tests and experiments have been designed in order to match the benchmarks listed in Deliverable D2.3. The deliverable is completed by a set of videos showing the experimental activities.

This project is funded by the European Union G.A No 101016906

Table of Content

1 Overview 5
2 Field Validation of the Robot Manipulation Functionalities 6
3 Field Validation of the Agronomic-Oriented Perception 19
3.1 Agronomic target detection and localization 19
3.2 Agronomic quality assessment 23
4 Field Validation of the Human-Robot Interaction 25
4.1 Evaluating the effectiveness of different means of communication 25
4.1.1 Participant Demographics 25
4.1.2 Virtual Reality 26
4.1.3 Tasks 27
4.1.4 Measures and Instruments 27
4.1.5 Usability 28
4.1.6 Competence & Discomfort 29
4.1.7 Trust 29
4.1.8 Results 29
4.1.9 Conclusions 33
4.2 Human-Robot Awareness 34
4.2.1 Evaluating Process Extraction 34
4.2.2 Evaluating Process Awareness 35
4.2.3 Experiments 36
5 Conclusions 39
6 List of attached videos 40
7 References 41
8 Appendix: Harvesting benchmark 42

This project is funded by the European Union G.A No 101016906

Abbreviations and Acronyms

ANOVA	Analysis of Variance
AUS	Alternative Usability Scale
BPM	Business Process Management
CNN	Convolutional Neural Network
DX.Y	Deliverable X.Y
HQP	Hierarchical-Quadratic-Programming
HRI	Human-Robot Interaction
HSD	Honestly Significant Difference
LLM	Large Language Model
LTL	Linear Temporal Logic
MAE	Mean Average Error
MAP	Mean Average Precision
PDDL	Planning Domain Definition Language
RER	Referring Expressions Recognition
RoSAS	Robotic Social Attributes Scale
SMEQ	Subjective Mental Effort Questionnaire
VR	Virtual Reality

This project is funded by the European Union
G.A No 101016906

1 Overview

This deliverable describes the validation of the Human-Robot Interaction (HRI) functionalities during harvesting and pruning as per Tasks 9.2, 9.3 and 9.4. The validation has been conducted in preparation of the final demo of the CANOPIES project.

The final demo shows the farming and logistic robots involved in a series of activities related to harvesting. These activities require the coordination of multiple functions including those related to HRI, perception, and manipulation.

Deliverable 9.2 consists of the current report together with a set of videos (see Section 6) showing the experimental activities in the field.

The deliverable is organized in the following sections:

- Section 2 covers the field validation of the robot manipulation functionalities (Task 9.2)
- Section 3 covers the field validation of the agronomic-oriented perception (Task 9.3)
- Section 4 covers the field validation of the Human-Robot-Interaction (Task 9.4)
- Section 5 concludes the report
- Section 6 describes the videos included in the folder mentioned above, linking them to the different tasks involved
- Section 7 includes relevant references

We designed tests and experiments to match the benchmarks defined in Deliverable D2.3. Each section contains for each identified benchmark (highlighted in bold italic), the evaluation activity conducted on field and/or in the virtual environment.

This project is funded by the European Union G.A No 101016906

2 Field Validation of the Robot Manipulation Functionalities

"Benchmarks for basic dual-arm functionalities. Capability of planning cooperative trajectories without violating kinematic and dynamic limits. Capability of avoiding kinematic singularities."

Here, it is shown how the control algorithm designed for planning the motion of the arm can handle robot kinematic and dynamic limits while avoiding singularities. The adopted algorithms are described in detail in D5.1 (*Cooperative arms and mobile base functionalities*) and are based on the Hierarchical-Quadratic-Programming framework which guarantees constraints' fulfillment by design if constraints at hand are at the highest priority within the hierarchy (this is the case in our design). Regarding kinematics limits, a sinusoidal reference trajectory with the aim of overcoming joint position and velocity limits is generated. Figure 1 in the top part reports the time evolution of the normalized joint positions (solid lines) with respect to the corresponding normalized minimum and maximum threshold values (dashed red lines) in the bottom part, instead, reports the time evolution of joint velocities (solid lines) with respect to the corresponding minimum and maximum threshold values (dashed red lines. Half of the limits are considered to maximally stress the algorithm). By observing these figures, it is possible to notice that throughout the entire experiment, both the joint positions and velocity limits are successfully respected in 100% of the cases without overshooting even in very demanding trajectories.

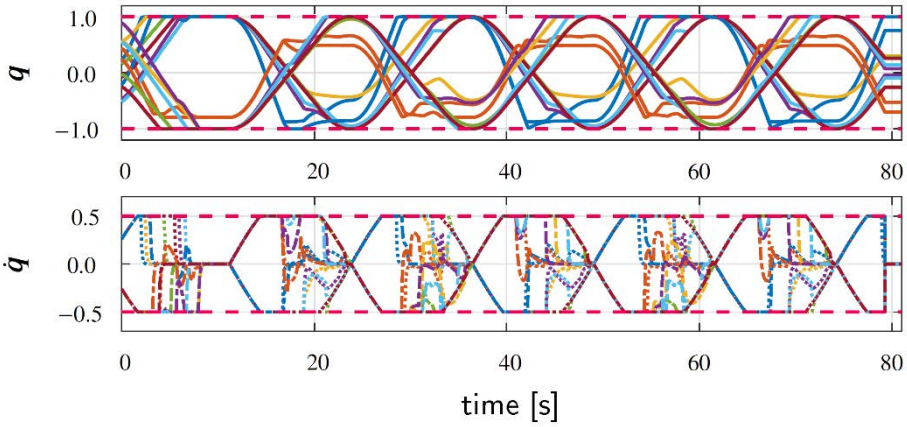


Figure 1. From the Top to the Bottom, it is displayed the time evolution of the normalized joint positions for left (dotted lines) and right (dot-dashed lines) arms and velocity and normalized imposed minimum and maximum thresholds (red-dashed lines).

Regarding the capability of the algorithm to handle kinematic singularities, a periodic motion of the end-effector aimed at completely stretching the arms in front of the robot (elbow and wrist singularities) is generated as shown in Figure 2.

This project is funded by the European Union G.A No 101016906

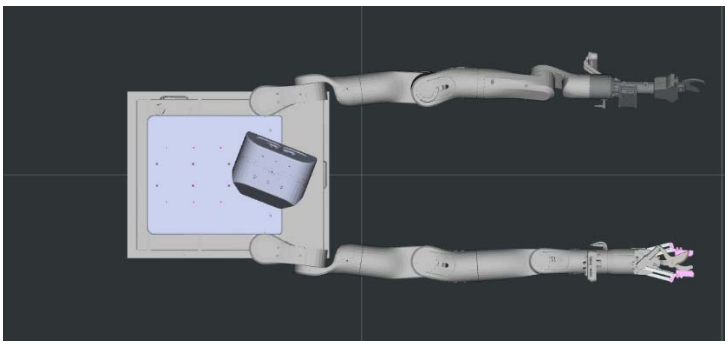


Figure 2. Desired configuration of the arms in the kinematic singularities stress test.

It is well-known that at algorithm singularities (theoretically) infinite velocities are generated since the smallest eigenvalue of the task Jacobian is singular, and a damping mechanism must be put in place to increase this eigenvalue. Figure 3 shows the time evolution of the joint positions (top plot) and of the joint velocities (middle plot) together with the smallest eigenvalue of the task Jacobian (bottom plot). In detail, it can be observed that the mechanism implemented within the HQP framework to handle kinematic singularities allows to keep the joints' velocities within provided limits even when the smallest eigenvalue is zero (worst case).

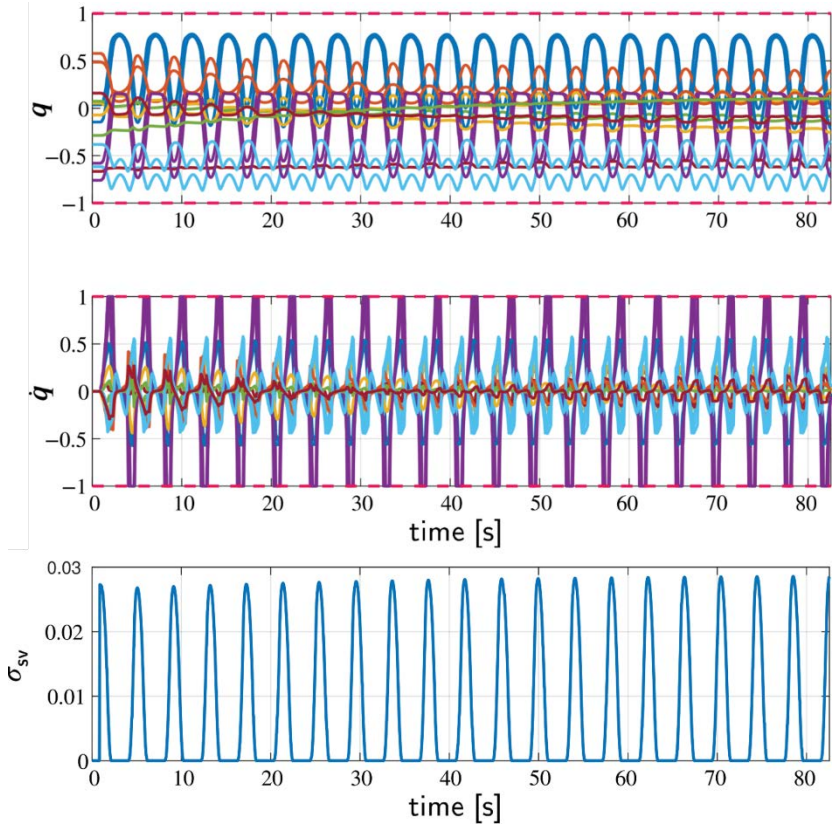


Figure 3. Joints' positions (Top), joints' velocities (middle), minimum singular value (bottom) of the task Jacobian matrix over time during the experiment.

"Capability of performing cooperative manipulation without collisions (given perception) with external objects and/or self-collisions."

Referring to the HQP control algorithm detailed in Deliverable 5.1, we show here the results of a test aimed at validating the safety features implemented to give the requested capability to the robot. The experiment has been carried out by using a hierarchy with three priority levels and five tasks, in detail, four safety tasks and an operational task. Since the aim of this experiment is to validate the safety tasks, we structured a lower-level (third level in the hierarchy) task that generates predefined sinusoidal velocities at each joint of both the arms, intentionally chosen to violate the constraints imposed by the higher-priority safety tasks.

The overall hierarchy is then the following:

1. Joint position and velocity limits
2. Virtual wall avoidance
3. Self-collision avoidance
4. Predefined sinusoidal joint motion

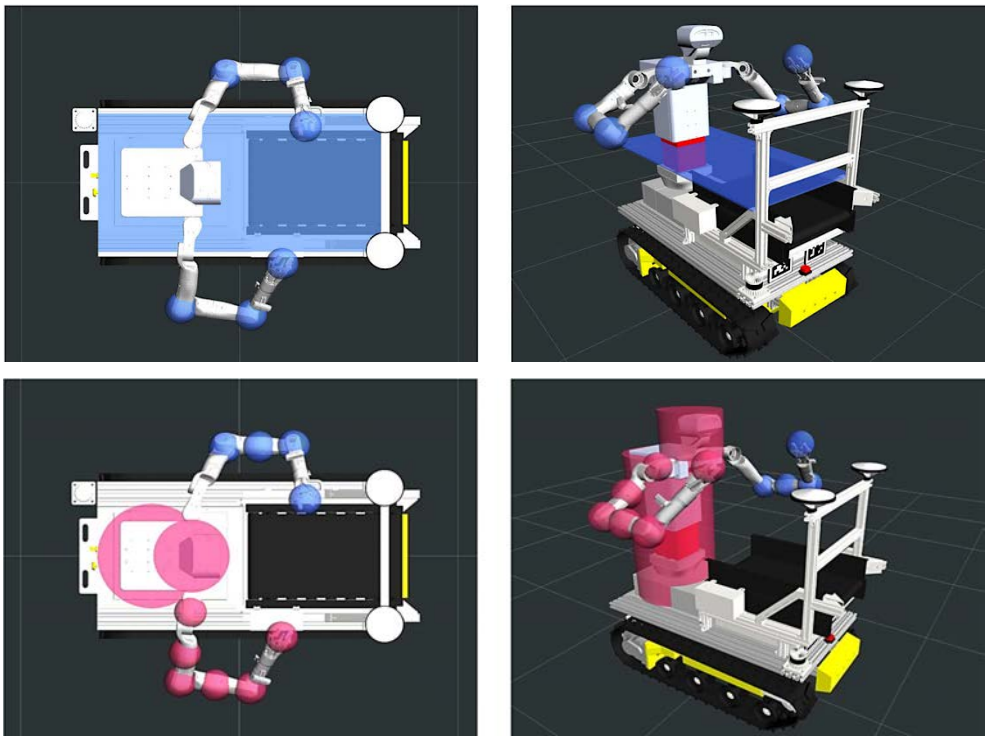


Figure 4. Graphical representation of the virtual wall task and the self-collision tasks

As anticipated above, the joint position limits have been chosen to mirror the actual limits of the actuators, while the maximum and minimum joint velocities have been intentionally limited at plus/minus 0.5 rad/s to activate the corresponding constraints.

The virtual wall task has been used to avoid collisions between the manipulators and the mobile base. Figure 4 contains a representation of the virtual wall with a top-down and a side view of the robot. More in detail, the blue surface represents the virtual wall itself, while the blue spheres represent the points of the manipulator structure subject to the constraint of keeping the distance from the virtual wall greater than the given threshold. For this experiment, the minimum distance to keep has been set at 0.3m.

The self-collision avoidance task has been used to avoid the occurrence of collisions between the manipulators themselves and between the manipulators and the other components of the bimanual robotic system, i.e. the torso and the head. Figure 4 shows representative points interested by the self-collision avoidance tasks. In detail, the images show the robot from different points of view, with interested points encapsulated within spheres and cylinders. In particular, the spheres highlighted in blue indicate the components whose position is conditioned by the constraints related to the task. The spheres and cylinders highlighted in red instead, in addition to indicating the components from which the ones encapsulated within the blue spheres must be kept at a distance, denoting the regions in which the thresholds associated with the task are not observed.

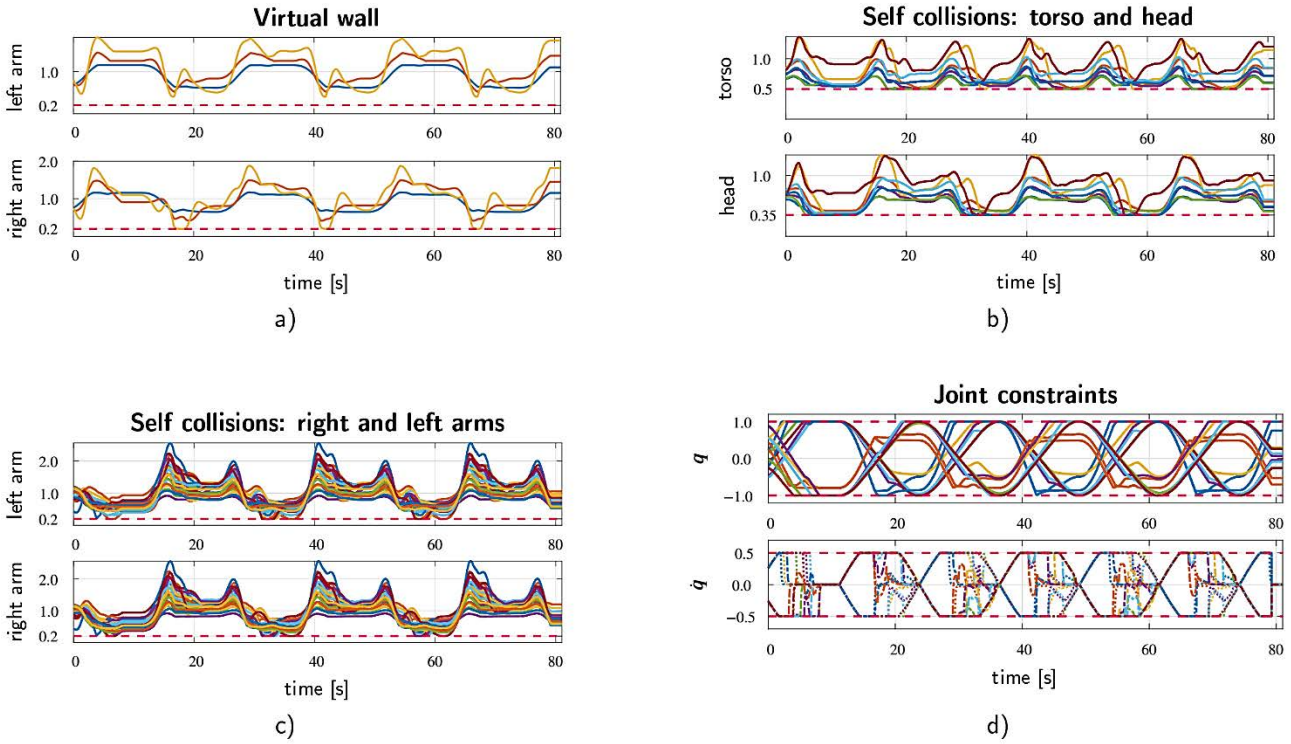


Figure 5. Results obtained during the experiment for validating the safety tasks. All the constraints are satisfied during the motion of the two manipulators.

Figure 5 shows the results obtained from this experiment. In detail, Figure 5.a) - Figure 5.c) show the time evolution of the distances related to the tasks of self-collision avoidance and virtual walls (solid blue lines) with respect to the corresponding threshold values (dashed red lines). Figure 5.d), instead, in the top part reports the time evolution of the normalized joint positions (solid lines) with respect to the corresponding normalized minimum and maximum threshold values (dashed red lines) and in the bottom part, instead, reports the time evolution of joint velocities (solid lines) with respect to the corresponding minimum and maximum threshold values (dashed red lines). By observing these figures, it is possible to notice that throughout the entire experiment, both the joint positions and velocity limits are successfully respected, and no one of the distances related to the tasks of self-collision avoidance and virtual walls takes values below the corresponding threshold.

This project is funded by the European Union G.A No 101016906

"Cluster collection in single arm or dual arm mode with a success rate of at least 70%."

Cluster collection occurs according to the procedure depicted in

Figure 6. The perception system provides the results of the detection and localization process (D4.1-Deliverable Agronomic target modelling, detection, and localization) to the control architecture, which then guides, the robot to bring its right harvesting tool to the pre-grasping position (

Figure 6. Top Right), i.e., at a predefined distance from one of the bunches whose peduncle has successfully been recognized. Then, the robot brings the tool to the grasping position and completes



the harvesting procedure performing the grasp

Figure 6. Top Left: image of the robot with both the harvesting tools in the home configuration. Top Right: image of the robot with the left harvesting tool in the so-called pre-grasping position. Bottom left: image of the robot with the left harvesting tool in the so-called pre-release position. Bottom right: image of the robot with the left harvesting tool in the so-called release position.

and the cut of the peduncle. Once the grape has been harvested, the robot brings the grape first to the pre-release position (

Figure 6. Bottom Left), i.e., at a predefined distance from the box where the grape must be placed, and then to the release position (

Figure 6. Bottom Right). Once the grape has been released, the robot brings the tool back to the initial configuration (

Figure 6. Top Left). Notably, in conditions as the ones represented in Figure 7 the farming robot is able to harvest. In this conditions, the grape is placed within the workspace of the robot (the control algorithm automatically computes the position of the platform to reach in order to have the grape in the robot's workspace, given the location of the grape computed by the perception system) and no significant obstacles (e.g., canes and iron wires) are close to the peduncle to grasp.

This project is funded by the European Union
G.A No 101016906



Figure 7. Typical configuration of the robot in a successful bunch harvesting.

Figure 8 shows, instead, an example of failing harvesting. In the right image, the harvesting gripper hits the berries because of a wrong estimation of the location of the gripper instead of reaching the corresponding peduncle.



Figure 8. An example of failing harvesting. The manipulator hits the berries instead of reaching the peduncle.

A complete set of images, showing the different cases is provided in Section 8. The reported cases show a success rate of about 79% which is greater than the benchmark (70%).

Regarding the dual-arm cooperative harvesting, in Figure 9, a typical sequence of dual-arm grape harvesting in partially realistic scenarios and during a winter campaign (winter 2023) with a dummy grape is shown. In the top left image, the arms are in the initial configuration in front of the robot and on top of the release box. In the top right image, the arms are in the pre-grasping position close to the bunch as located by the perception system; then, in the bottom left image, the right gripper holds the bunch cut by the left end-effector. Finally, in the bottom right image, the right arm releases the bunch in the collecting box. However, this was a temporary setup that has been replaced with the final setting that involves two different end effectors for harvesting and pruning, respectively.



Figure 9. Snapshots of the farming robot in a cooperative harvesting operation. Top Left: arms in the initial configuration. Top Right: arms in pre-grasping position. Bottom Left: arms in post-grasping position. Bottom Right: right arm in releasing grape configuration.

"Cluster placing in single arm or dual arm mode with a success rate of at least 90%."

Regarding the placing of collected grapes in the placing box (in Figure 10), the success rate of the operation is 100% since it is obtained by design. In detail, because of the system design (described, in particular, in D2.1- *Requirements, Specifications and Benchmarks* and D2.2- *Specifications and KPIs for the two robot prototypes*) the release box (in green in Figure 10) is within the robot workspace and the release locations offline optimized in order to be reachable by the robot arm.



Figure 10. The farming robot with the placing box (green box) on the BEM mechanism.

"Human safety guarantee via definition of a safety field accounting for human and robot relative position and velocity and for which a minimum threshold is set. Interaction is considered safe if the safety field is kept above the threshold."

This project is funded by the European Union G.A No 101016906

To respect this benchmark, we designed on-field experiments based on the Safety Planner detailed in Deliverable D5.2 (D5.2-*Mobile arms safety functionalities*). More in detail, in the experiments, a human operator shares the workspace with the robot while it performs the harvesting operation, as in Figure 11.



Figure 11. Pictures of two experiments performed on the field. Preliminary (left) and final (right) trials.

The pictures show two experiments in different environmental conditions conducted to validate the designed human-aware safety planner (a preliminary experiment on the left and the final on the right). Initially, the robot follows a nominal trajectory until the human operator gets closer, triggering the activation of the constraint regarding the minimum safety field value. The results reported in Figure 12 are related to the final experiment (on the right in Figure 11). Results are anyway consistent between the different experiments conducted as the proposed methodology ensure safety by design. More in detail, Figure 12.a) shows the evolution of the safety field value over time. The human operator gets close to the robot twice during the experiment. Specifically, around t=10s, the human approaches the robot and the safety field value reach proximity of the minimum threshold of 0.75. This results in a deviation from the nominal trajectory (Figure 12.b)-Figure 12.d)) for about 10s.

This project is funded by the European Union G.A No 101016906

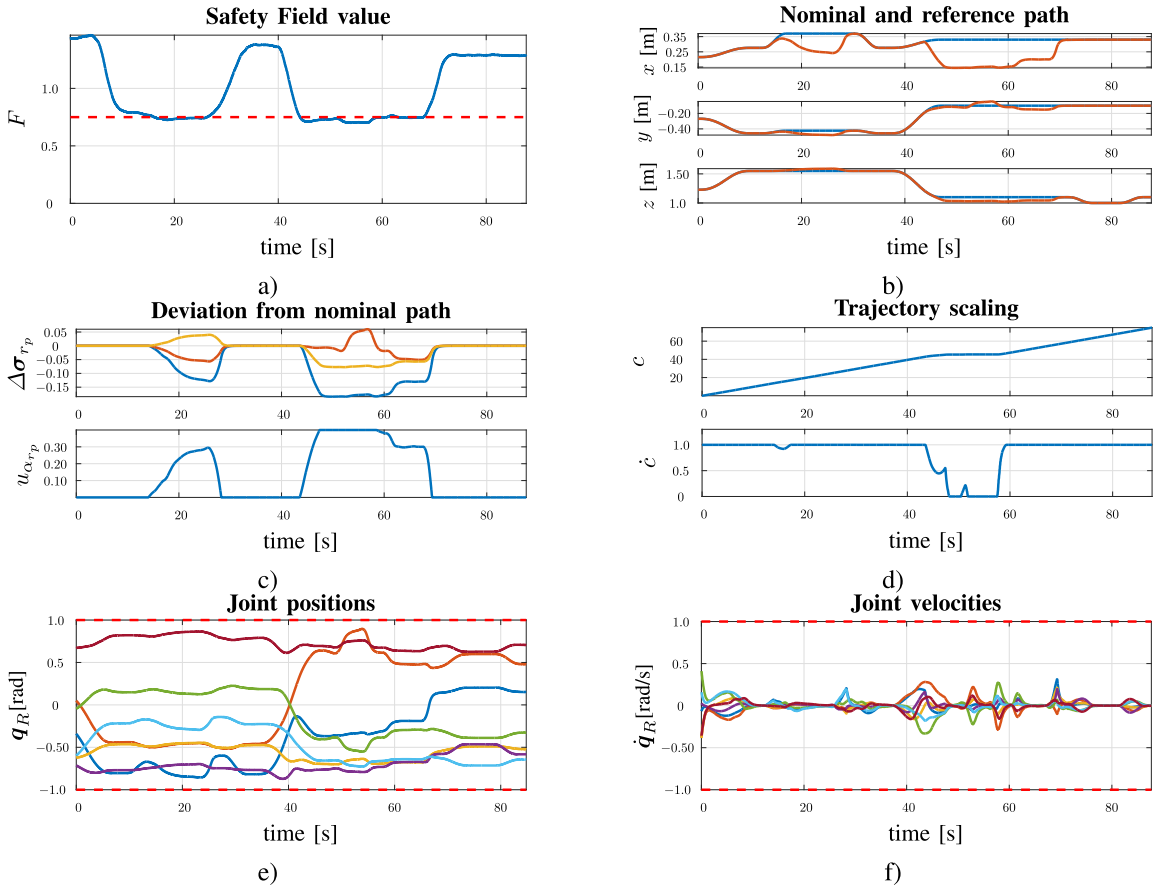


Figure 12. Field experiment. a) Safety field value over time and imposed minimum threshold (red-dashed horizontal line); b) Nominal (blue) and safe (red) trajectories for the absolute frame. The nominal trajectory is composed of a sequence of waypoints suitable for performing a harvesting operation. The proposed Safety Planner changes the nominal path to keep the safety value above the imposed minimum threshold; c) Deviation from the nominal path. d) Parameter c and virtual input. e) Normalized joint position with the imposed thresholds. f) Normalized joint velocities with the imposed thresholds.

Then, the human operator steps back, and the robot resumes the tracking of the nominal trajectory without modifications. Around t=50s, the operator gets close again to the robot, leading to a further phase, with a duration of about 20s, where the path is modified. It is worth noticing that in this phase, the safety field value goes below the minimum threshold because both the virtual inputs have reached their respective maximum and minimum values. In this condition, the algorithm maximizes the safety field value, but there is no guarantee of respecting the minimum threshold. This happens since, being the safety based on the distance between the human and the robot, if the human keeps moving in the direction of the robot, the latter should keep moving in the opposite direction to keep the distance (and then, the value of the safety field) above the limit value. However, if the mobile base is kept still (as in this case given the agronomic task at hand) the safety distance cannot be kept above the threshold given the maximum allowed displacement from the nominal trajectory (by setting a maximum value for $u_{\alpha_{r_p}}$ -- 0.3 in this case as shown in Figure 12.c)), the limited workspace of the robot arms and the necessity of preventing robot self-collisions (between the right arm and the robot's torso in this case). However, it is worth noticing that, even in this situation, the safety of

This project is funded by the European Union G.A No 101016906

the human operator is still preserved, since the overall approach is designed such as $\dot{c} = 0$ in this case, meaning that the robot does not move at all (Figure 12.c) and Figure 12.d)) and Figure 12.e) and Figure 12.f) demonstrate that the joint limits are respected during the entire experiment. Finally, at t=69s, the human operator steps back, allowing the robot to finish the harvesting operation. Finally, since the overall approach is based on the adoption of the HQP framework described in D5.1 and D5.2 mentioned above, and since the safety of the human represent the highest priority task, the safety of the human is guaranteed by design according to the defined metrics.

"In tasks for which physical interaction is allowed, capability to adjust the arm and end-effector configurations according to human exerted wrench with a success rate of at least 90% and without instability phenomena."

To describe and assess the robot's physical interaction capabilities, the overall functional architecture in which it operates is first outlined. Referring to the shared control strategy detailed in Deliverable D5.3 (D5.3-*Mobile arms physical interaction*), a high-level layer on top of the HQP framework is in charge of changing the robot autonomy according to the context and in order to increase the effectiveness of the harvesting operation. Two possible control modes are envisioned: *autonomous* mode and *guided* mode. In autonomous mode, the robot autonomously performs the needed end-effector motions to perform the entire harvesting procedure in case perception succeeds in finding the peduncle. In guided mode, part of the harvesting process is managed by the human operator in case of failures of the robotic system (e.g., mainly because grapes might be occluded by leaves). In detail, at the beginning the robot is under autonomous control mode and runs the perception software making use of the RGB-D sensor of the head. In cases in which bunches are detected and localized by the perception system and peduncles are not clearly visible, the perception system can only provide a rough estimate of the position of the peduncle and give this information to the robot to move the arm closer and take a new image from the arm camera. If the peduncle is not seen also from the arm camera, then the cutting point cannot be calculated, and physical interaction is required (guided mode). In this case, the human operator guides the end-effector towards the point to cut by physically moving the end-effector. Then, the robot control mode switches back to autonomous, and it continues the remaining steps of the harvesting procedure on its own. An example of obtained results in this scenario are shown in Figure 13, where Figure 13 (top) shows the interaction wrench of the right end-effector during the harvesting experiment, Figure 13 (middle) reports the normalized joint positions and velocities and the corresponding thresholds, and Figure 13 (bottom) depicts the x, y, and z coordinates of the actual position of the end-effector and the corresponding desired values.

This project is funded by the European Union G.A No 101016906

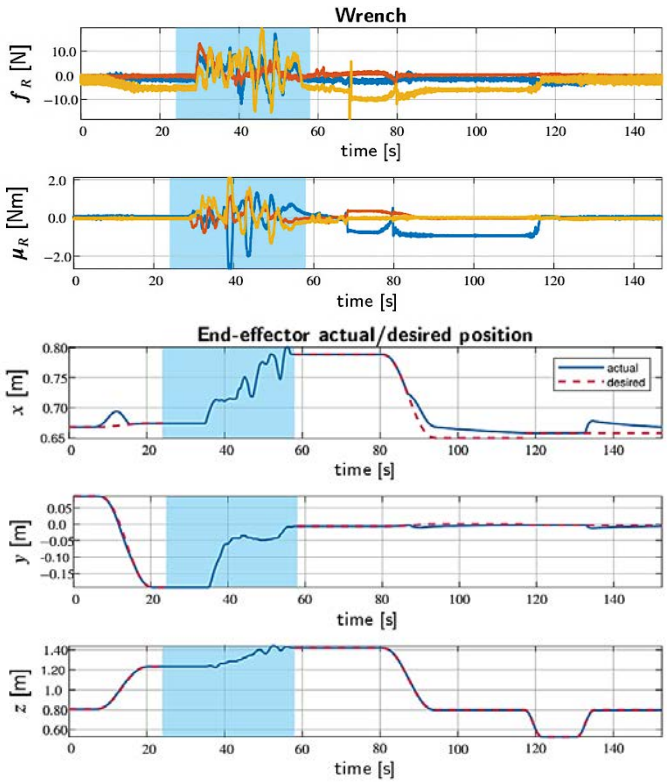


Figure 13. Top. Interaction wrench applied by the human operator on the right end-effector throughout the harvesting experiment. Bottom. Time evolution of the end-effector trajectory accommodating the human wrench in the interaction phase.

The cyan background of the top and bottom plots highlights the phase in which the robot is in *Hand-Guiding Mode*, while the white background indicates that it is in autonomous mode. More in detail, from t=0s to t=20s the robot moves the right end-effector in a position at a predefined distance from one of the bunches and makes a new estimation attempt. As mentioned before, even in this case, the peduncle of the target bunch was not recognized and, at this point, the robot asks for the human operator's assistance, and its control mode switches to guided and remains in this state until about t=58s. In this phase, the admittance parameters are changed according to this mode, allowing the human operator to grab the end-effector and manually place it on the point to cut. This is clear from Figure 13 (top), which shows large interaction forces and moments during this phase, and from Figure 13 (bottom), which does not show any desired trajectory given that the operator freely drives the end-effector by hand. It can be noticed that the end-effector moves according to the generated wrenches that do not reach critical values. Furthermore, no instability phenomena are visible for the whole duration of the interaction. This held true also in other tests on the field as well as all in laboratory tests (in this regard, no significant difference exists between laboratory and field tests since only force exerted by human operators and control algorithms affect the final results). Figure 14 shows some pictures of four experiments conducted on the field for validating the shared control architecture and in which the human physically interacts with the robot to bring the end-effector to the correct location to harvest the bunch. When the operator releases the end-effector, the robot detects this condition by comparing the norm of the interaction forces and torques with given thresholds and the robot switches back to autonomous mode. From that point on, the robot autonomously continues the harvesting procedures with the admittance parameters related to the

autonomous control mode. It is important to highlight that the capability of the robot to accommodate human motion is obtained by design through the designed force strategy (see D9.2) according to which the robot fully follows the desired trajectory commanded by the human that can bring the robot in any desired location. In this perspective, a percentage of 100% of success can be considered for the given benchmark.



Figure 14. Pictures of two experiments performed on the field relative to human-robot physical interaction. In the bottom pictures the second human is also exploiting the designed Human-Machine-Interface.

"In tasks with both human interaction and moving mobile base, the latter is within the conditions detailed in objective 3.4.1, i.e., keeping a spatial robustness (see brief definition in 3.4.1) of 5% of the distance between plants according to D2.1, Table 1."

This aspect has been analyzed in D9.3. In particular, in Section 8. For the sake of simplicity, we report in the following what written in that deliverable:

"Regarding the spatial robustness, the LTL planner works through the discretization of the workspace where some cells were reserved for the poles of the vineyard which act as the obstacles/plants referred to in the 3.4 benchmarks. Our planner ensures that those pole cells are always avoided and tries to ensure that the robot goes through the center of the empty cells as shown in Fig. 29. This means that the planner, by design, ensures that there is always a safety distance equal to at least half the side length of the square cell and ideally equal to the side length of the square cell. Since the lower-level navigation and localization stack takes care of following these waypoints, we refer to the benchmarks ``1.1 Robots navigation" for more details."

This project is funded by the European Union G.A No 101016906

"Benchmarks for evaluating the designed end effector (i) Capability of holding 2kg payload, (ii) Capability of cutting and hold 20mm diameter branches, (iii) Capability of cutting 50mm branches."

Concerning point (i), we conducted several tests by loading the robotic arm with progressively increasing weights: 0.5 kg, 1.0 kg, up to a maximum of 2.5 kg. During all the tests, the robot was able to support the load and maintain the required position. In particular, when the arm was fully extended, it was necessary to increase the current limits on the joints most affected by the weight to compensate for the additional load. With these adjustments, the robot demonstrated its ability to securely maintain its position even under maximum load.

Concerning point (ii) and (iii) the pruning gripper has successfully cut a double branch exceeding 50mm in diameter. Also, the harvesting gripper was able to cut and hold a 20mm diameter branch.

This project is funded by the European Union G.A No 101016906

3 Field Validation of the Agronomic-Oriented Perception

3.1 Agronomic target detection and localization

"Benchmarks for cluster detection and localization for harvesting. Test data will be collected and labelled to test the accuracy of the algorithms. The algorithms will be compared with the state of the art of cluster detection and localization in harvesting, for example FASTER R CNN. The following test will be done:

- ***Test of accuracy in cluster detection with a success rate of at least 70%.***
- ***Test of accuracy in cluster localization with a success rate of at least 70%."***

The benchmarks were done with a dataset recorded in September 2021, of grape bunches and peduncles that were collected in the vineyard of the Corsira Agicultural Cooperative Society (Aprilia, Italy). The images were acquired on different days and at different times, from 9am to 5pm. The images were also taken at different distances from the bunches, ranging from 30cm to 2m.

Ten sequences of images were selected in which the peduncles were fully or partially visible. The CANOPIES Grape Bunch and Peduncle dataset [7] contains a total of 810 RGB images and with their correspondent binary masks for grapes and peduncles. This dataset is in the Zenodo repository.

The results obtained for cluster and peduncle detection and localization using a confidence score threshold of 0.8 and IoU threshold of 0.3, were the following:

- Using the mean Average Precision (mAP) metric for RGB-D images, our method obtained a mAP = 0.767 which outperforms the state-of-the-art methods (see article [6]) that use RGB data (their results are lower, mAP = 0.725).
- Regarding depth estimation, our method achieves a mean squared error of 2.66cm within a distance of 1m.

The same method was applied to the standard WGISD dataset of clusters, and the results were the following:

- Using the mean Average Precision (mAP) metric for RGB-D images, our method obtained a mAP = 0.949 which outperforms the state-of-the-art methods that use RGB data (their results are lower, mAP = 0.891).

For more details refer to our published article [6].

"Benchmarks for canes detection in pruning. Test data will be collected and labelled to test the accuracy of the algorithms. The algorithms will be compared with the state of the art for canes detection in pruning. The following test will be done:

- ***Test of accuracy in canes detection with a success rate of at least 70%.***
- ***Test of accuracy in canes localization with a success rate of at least 70%."***

This project is funded by the European Union G.A No 101016906

Benchmarks for canes and buds' selection for pruning. Test data will be collected and labelled to test the accuracy of the algorithms. The algorithms will be compared with the state of the art for canes and buds' selection in pruning. The following test will be done:

- ***Test of accuracy in selection of canes and buds with at least 50% compliance with respect to the cutting suggestions provided by an expert."***

After performing an assessment of the current state-of-the-art solutions regarding robotic pruning, such as [1][2][3][4], their implementation is not feasible in a setting such as the vineyards in the Corsira Agicultural Cooperative Society (Aprilia, Italy). The sheer size of the vines and their consistent intertwining do not allow for an adequate reconstruction without multiple images collected from various positions including aerial ones, which is not possible given the available hardware.

A first iteration of the perception pipeline was tested on a total of 7 video sequences containing a total of 6372 images from the Aprilia field recorded from January to February 2023 and stored in rosbag files (this dataset will be available when the article that use it is published). An experimental validation testing phase with the final version of the perception pipeline deployed on the robotic platform alongside the manipulation and navigation software was performed during the week of 3rd of March 2024 to the 8th of March 2024 in the Corsira Agricultural Cooperative Society, with the aim of testing the correct functionality of the perception solution under real life conditions. A fine tuning and quantitative validation testing were conducted during the month of June 2024 on a mock-up built in the UPC that reproduces the vineyard roof setup.

The experiments procedure regarding the perception pipeline was the following:

- The agronomist expert would point to the branch section where the pruning must be performed.
- The perception pipeline detects the hand of the agronomist and defines the region of the scene where the final pruning pose must be identified.
- The perception pipeline identifies the different branches appearing in said region and singles out the objective branch where the pruning must be performed.
- The gem buds of the objective branch are detected, and the two that delimit the section where the pruning must be performed.
- A pruning pose located in the middle of said section is extracted and communicated to the robot.

The experiments result of the objective branch identification algorithm based on the use of Random Sample Consensus gave error metrics reported in Table 1.

Mean error of the detected position of the branch objective with respect its real position	0.0104 ± 0.0094 m
Mean error of the detected orientation of the branch objective with respect its real orientation	16.21 ± 16.33º

Table 1. Error metrics.

Regarding the success rate in identifying the correct branch, it ended with 126/141 correct detections for the average case, with 145/182 when accounting all tested cases, including "too complex/hard to find" cases, i.e. scenes with numerous branches and exogenous objects such as wires or poles appearing in the scene.

In Figure 15 and Figure 16, there can be seen the process of detection of the point to prune and the branch pruned.



Figure 15. (a) A person shows where the robot must cut; (b) the robot approaches to cut the branch; (c) the branch is cut.



Figure 16. (a) A person shows where the robot must cut; (b) the robot approaches to cut the branch; (c) the branch is cut

Regarding the success of canes localization for cutting, you can see that in the experiments, all points for cutting were well detected and the branch was cut that means that they were located the 100% of the time.

The experiments result of the gem bud detection algorithm based on the use of convolutional neural networks presented the metrics reported in Table 2.

Number of experiments	100
Number of gem buds to detect	639
Number of detected gem buds	524
Number of missed buds	115
Number of false positives	207
Precision	0.72
Recall	0.82
F1-Score	0.76

Table 2. Results for the gem bud detection algorithm.

It is worth noting that most of the false detections during the experiments were located at the border of the images depicting the region of the scene where the final pruning pose must be identified. These errors in the detection are, while not trivial, not especially relevant, as in the immense majority of experiments there were more actual gems detected in zones closer to the center of the image, which is the position used to select the gem bud pair to be used in the identification of the section where the pruning must be performed, thus making this false detection around the limits of the image very unlikely to be selected as components of said pair. By subtracting these false detections in the images border, the metrics are reported in Table 3.

Number of experiments	100
Number of gem buds to detect	639
Number of detected gem buds	524
Number of missed buds	115
Number of false positives	52
Precision	0.91
Recall	0.82
F1-Score	0.86

Table 3. Results for the gem bud detection algorithm without false positives.

As it can be seen in the results, for the branch detection the results were 89% correct detections for the average case, with 79,7% when accounting all tested cases, including "too complex/hard to find" cases.

For the gem bud detection, the F1-score was 76% in general, and eliminating the false detections at the border of the images, the result was 86%.

In both cases, we have fulfilled the benchmarks that were established for this project.

3.2 Agronomic quality assessment

"Benchmarks for Quality Assessment for selective Harvesting. Test data will be collected and labelled to test the accuracy of the algorithms and quality assessment procedures. More in detail, the algorithms should be able to have high (close to 80% of human performances, on average) accuracy on selection of grape bunches to harvest according to specific quality parameters, as described in D2.1. With this respect, the algorithms should pass the following tests:

- ***Berry size estimation: achieve a mean absolute error (MAE) less than 30% of berry size on benchmark data.***
- ***Color classification: achieve at least 80% of accuracy on grape bunch color classification on benchmark data.***

This project is funded by the European Union G.A No 101016906

- *Sugar content estimation: achieve a MAE less than 5 °Brix with respect to the measured value of °Brix sugar level using visual sensors on benchmark data. - Density estimation: achieve at least 80% of accuracy on bunch density estimation on benchmark data. - Defects estimation: achieve at least 80% of accuracy on defects and sickness identification on benchmark data."*

An extensive evaluation of the agronomic quality assessment, respecting the defined KPIs, has been provided in Deliverable D4.2. The report gives technical information about the problem of quality estimation of table grapes using RGB-D inputs. Data collection campaigns during the two summer seasons of 2021 and 2022 have been described, together with the details of the post-processing and labelling of training and validation data. For the sake of simplicity, we report in the following a summary of the results.

Methods for the estimation of all four quality parameters have been presented, and all of them have been validated on datasets collected in the experimental field. The results achieved are the following:

- Anomaly detection for harvesting (binary decision): 87% precision
- Colour for harvesting (binary decision): 84% accuracy
- Brix° estimation: 1.08 MAE
- Berry size MAE: 3.7% average error with 16%

In all cases the performance target KPIs are outperformed.

This project is funded by the European Union G.A No 101016906

4 Field Validation of the Human-Robot Interaction

4.1 Evaluating the effectiveness of different means of communication

"Benchmarks for evaluating the effectiveness of the different means of communication. The effectiveness of the different means of communication will be evaluated by first asking expert users to evaluate communication acts in the virtual environment and to fill questionnaires based on Likert scales. As WP6 will define tens of different communication acts, a subset of these will be selected for the evaluation. The outcome of this evaluation can then be used to modify the system accordingly. An evaluation can then be performed on the field. A first goal will be to have a percentage of correctly detected acts between 70% and 100% according to the cruciality of the communication act itself. Questionnaires instead must assess the appropriateness of the communication modality with respect to the gravity of the specific situation. In addition, the evaluation must ensure that the communication modalities can scale up when multiple humans and robots are involved in the tasks. The questionnaires will be defined taking inspiration from the Robotic Social Attributes Scale (RoSAS). A second goal will be then to ensure that, for each selected communication act, the average score of reported answer is at least 20% above normal, i.e., if 3 is the normal score, 3.6 should be the minimum average scores reported by users."

In this section, we evaluate the explicit communication modalities as defined in D6.2.

The primary motivation for conducting a multimodal interaction study stems from the challenges encountered in outdoor environments. As discussed in the introduction, these settings are frequently noisy, posing difficulties for a single modality to provide all the necessary information for effective robot performance. This study investigates the impact of multimodal input, precisely the combination of speech and gestures, on user experience. Each modality has been individually assessed to comprehensively evaluate this impact, resulting in three distinct input settings: speech, gestures, and multimodal interactions. During our experiments, if speech input proves inadequate, the robot prompts the user for additional details using either speech or gestures. If speech fails to provide the required information, gestures take precedence. This approach ensures the robot can accurately interpret commands and execute tasks effectively, capitalizing on the strengths of both modalities to tackle the complexities of outdoor settings.

The gestures designed for the CANOPIES project were carefully curated to ensure they do not overlap with any natural spoken gestures, making it challenging for participants to learn how to interact with the robot using these gesture-based modalities. If a participant were first to experience the gesture-only modality and then the multimodal setting (or vice versa), they would have an advantage in the second experiment, which could undermine the study's validity. A between-subjects design was employed to prevent bias, ensuring participants engaged with only one modality during the study. This design decision aims to maintain the integrity and validity of the research findings.

4.1.1 Participant Demographics

To conduct the experiments, we recruited 90 participants from the Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti (DIAG) of Sapienza University and the Istituto

Tecnico Agrario Emilio Sereni in Rome. All data collected was anonymous, with no personal identifiers. Participants had no speech impairments or physical disabilities, enabling them to participate in any of the three modalities without constraints.

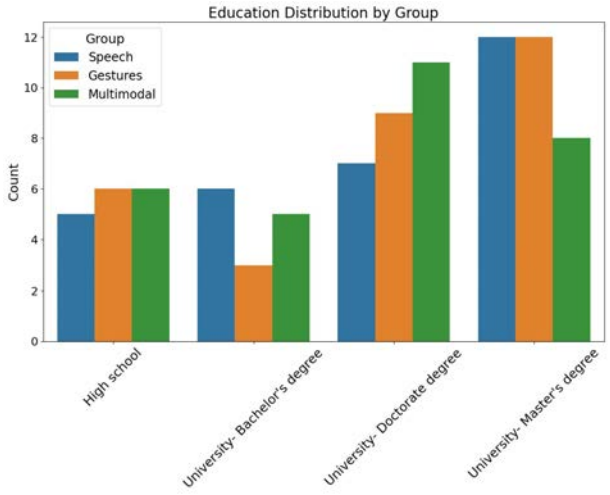


Figure 17. Distribution of participants' level of education per group.

The participants were evenly divided into three groups, with 30 individuals per group (see Figure 17). Among the 90 participants, 66 identified as male, 22 as female, 1 as 'other', and 1 preferred not to disclose their gender. Educational backgrounds varied: 11 participants were in high school, 6 had completed high school, 14 were enrolled in a bachelor's program, 32 were pursuing a master's degree, and 27 were either pursuing or had completed a doctorate. The average age of the participants was 27.49 years. Table 4 presents the mean age and gender distribution of each group, while Figure 17 illustrates the education level distribution per group.

Regarding prior experiences, 52 participants had interacted with virtual reality (VR) before. On a 7-point Likert scale (ranging from 'never' to 'daily'), participants rated their experiences with robots in VR at an average of 3.18, indicating infrequent interaction with simulated robots. Similarly, their experience with real robots averaged 3.15 on the same scale. Lastly, when asked about their experience in vineyards, the average score was 3.08, suggesting limited familiarity with vineyard environments.

4.1.2 Virtual Reality

The proliferation of virtual reality (VR) technologies has opened up new avenues for studying human-robot interaction (HRI) in various contexts. The efficacy of utilizing VR for HRI research in agriculture has been demonstrated within the CANOPIES project. Moreover, the logistical challenges associated with transporting both the robot and participants to a vineyard environment have been addressed by conducting experiments within a simulated environment. In this simulated setting, participants' real actions, such as speech or gestures, directly influence the behavior of the simulated robot, allowing for a more controlled and efficient study of HRI interactions.

Input modality	Age	Male	Female	Other	Prefer not to say
Speech-only	25.60	20	10	0	0

Gestures-only	26.43	25	5	0	0
Multi-modal	30.43	21	7	1	1

Table 4. Experimental groups.

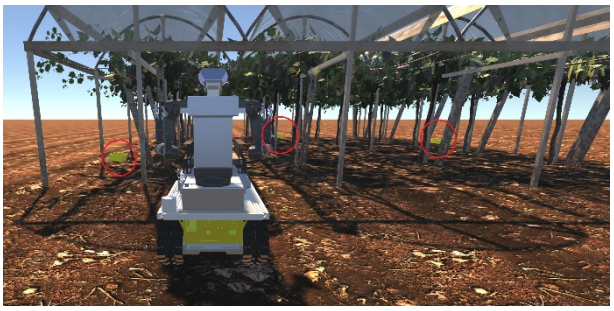


Figure 18. Participant's initial view of the simulation, each red circle depicts the placement of a box and was not part of the simulation.

4.1.3 Tasks

All participants, regardless of group, were tasked with instructing the robot to complete specific actions. The tasks were designed to be identical across all groups and involved navigating the robot to three boxes placed in the field. The boxes were positioned to the right, in front of, and to the left of the robot's starting point (see Figure 18). Participants were required to navigate the robot to each box, and upon approaching each one, the robot would automatically collect it. Once all boxes were collected, participants had to direct the robot to harvest a grape bunch. Instructions were to be conveyed in the participant's assigned modality.

The instructions included six commands: ***move forward, move backward, turn left, turn right, stop, harvest***. Since the participants were Italian, the vocal commands for the speech and multimodal groups were given in Italian. However, the system could be set to other languages if necessary. The specific Italian commands used were: ***'vai avanti'*** (move forward), ***'vai indietro'*** (move backward), ***'gira a sinistra'*** (turn left), ***'gira a destra'*** (turn right), ***'fermati'*** (stop), and ***'raccogli'*** (harvest). The associated gestures for the gestures and multimodal groups are shown in Figure 19.

4.1.4 Measures and Instruments

A comprehensive questionnaire was administered to assess the participants' experiences after interacting with the robot to gauge key concepts: usability, trust, perceived competence, and discomfort. Since the participants were native Italian speakers with varying levels of English proficiency, the questionnaire was translated into Italian and reviewed by individuals proficient in both languages to ensure clarity and accuracy, thus mitigating potential language related ambiguities.



Figure 19. Gesture commands.

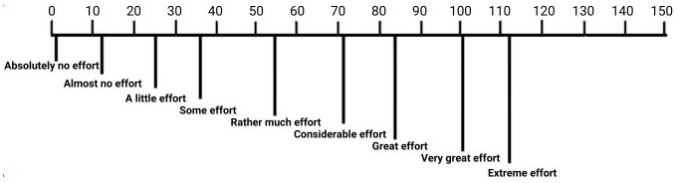


Figure 20. Subjective effort scale.

4.1.5 Usability

The Alternative Usability Scale (AUS) was employed to evaluate usability, which was extensively validated and researched. This questionnaire includes a single item for each usability component, rated on a 7-point Likert scale ranging from 'strongly disagree' to 'strongly agree.' This allows for the analysis of learnability, memorability, efficiency, satisfaction, and error recovery, which are the five components of usability defined by Nielsen's Five-Factor Model.

In addition to these subjective measures, the perceived effort expended during the experiment was also assessed. This included mental and physical effort, measured using the Subjective Mental Effort Questionnaire (SMEQ) and an adapted version of the SMEQ, respectively. The SMEQ asks users to rate the effort expended on the interaction on a scale from 0 to 150, as depicted in Figure 20.

Objective usability measures were also employed to assess the participants' performance. Task duration was used as a critical metric, measuring users' time to complete the tasks. This timing was used to evaluate relative user efficiency, comparing participants' task completion times with expert users.

Competence items	Discomfort items
Reliable	Scary
Competent	Dangerous

This project is funded by the European Union G.A No 101016906

Knowledgeable	Aggressive
Interactive	Strange
Responsive	Awful
Capable	Awkward

Table 5. RoSAS items.

4.1.6 Competence & Discomfort

Both competence and discomfort are social perceptions of robots that can be measured using the Robotic Social Attributes Scale (RoSAS), a standardized 18-item questionnaire designed to assess perceived competence, warmth, and discomfort. Given the CANOPIES robot's utilitarian nature, the warmth subscale was excluded from the questionnaire. A detailed breakdown of the items evaluated in the RoSAS subscale can be found in Table 5. Each item was evaluated by participants on a 7-point Likert scale to determine the extent to which each item accurately describes the robot.

4.1.7 Trust

In the experiment, trust was one of several measures assessed, necessitating a concise questionnaire to minimize user burden. This led to the selection of a 3-item trust questionnaire. Additionally, since one of the items overlapped with the Robotic Social Attributes Scale (RoSAS) reliability measure, only two items were included: 'The robot is trustworthy' and 'I can trust the information presented by this robot'. Both items were evaluated on a 7-point Likert scale, ranging from 'strongly disagree' to 'strongly agree'.

4.1.8 Results

To facilitate statistical analysis of mental effort and physical effort, the participants' inputs had to be transformed. The increments of 10 were categorized from 0 to 14 in ascending order (e.g.,0-10 was turned into 0, 11-20 was turned into 1, and so on). The error recovery item from the usability scale was reversed and then averaged with the other four usability items (memorability, learnability, efficiency, and satisfaction). Due to their diverse and informative nature, the items of usability were also analyzed individually. The two items of trust were averaged into a single trust measure. Additionally, the six items of perceived competence and the six items of perceived discomfort were averaged into two single measures.

To ensure the quality of the data, the resulting variables were checked for outliers within each group using the interquartile range (IQR) method. Thereafter, the data was scaled on a range of [0,1] for each variable, and outliers were imputed with the mean for each variable in their respective groups.

After the transformations were completed, the mean score of each variable was calculated for the three groups, which can be seen in Table 6.

In order to determine the most suitable test for comparing means, it is essential to assess the normality of the distributions. The Shapiro-Wilk test, recognized as one of the most robust formal normality tests, particularly for small sample sizes, was employed for this purpose. Using a significance level of $\alpha = 0.05$, the following variables within each group demonstrated normality:

- **Speech:** relative user efficiency, memorability, physical effort, and perceived competence.

This project is funded by the European Union G.A No 101016906

- **Gestures:** relative user efficiency, usability, efficiency, trust, and perceived competence.
- **Multimodal:** relative user efficiency, usability, error recovery, trust, and perceived competence.

Among these, ***relative user efficiency*** and ***perceived competence*** were the only variables that met the normality assumption across all three groups. Given that not all variables exhibited a normal distribution, a combination of methods was employed to compare differences between group means. For variables meeting the normality assumption in all groups, analysis of variance (ANOVA) was utilized, while the Kruskal-Wallis test was employed for variables not meeting this assumption. Despite only two variables meeting the criteria for ANOVA, this approach was justified due to its enhanced statistical power compared to the Kruskal-Wallis test.

Table 7 provides an overview of statistically significant differences observed across group means, with detailed explanations of each variable presented subsequently.

Measure	Speech	Gestures	Multimodal
↓ Relative user efficiency	0.21	0.46	0.50
↑ Usability	0.58	0.62	0.56
↑ Efficiency	0.62	0.70	0.59
↑ Satisfaction	0.78	0.77	0.77
↑ Memorability	1.00	0.84	0.69
↑ Learnability	0.92	0.91	0.80
↓ Error recovery	0.50	0.39	0.47
↓ Mental effort	0.20	0.20	0.30
↓ Physical effort	0.00	0.25	0.26
↑ Trust	0.72	0.68	0.63
↑ Perceived competence	0.67	0.72	0.66
↓ Perceived discomfort	0.15	0.09	0.17

Table 6. Average scores (scaled from 0 to 1). Measures marked with an ↓ indicate negative aspects, where lower scores are better.

Measure	Speech	Gestures	Multimodal
* Relative user efficiency	0.21	0.46	
* Relative user efficiency	0.21		0.50
Memorability	1.00	0.84	
Memorability	1.00		0.69
Learnability	0.92		0.80
* Mental effort	0.20		0.30
* Physical effort	0.00	0.25	
* Physical effort	0.00		0.26

Table 7. Significant differences. Measures marked with an asterisk indicate negative aspects, where lower scores are better.

Relative user efficiency: The differences in memorability between modalities (see Figure 21) revealed significant differences in relative user efficiency, with a p-value of 3.17×10^{-8} obtained from the ANOVA test, indicating statistical significance at $\alpha = 0.05$. However, the ANOVA test alone indicates only that there is a significant difference among at least two of the three different groups. To delve deeper into these discrepancies, Tukey's Honestly Significant Difference (HSD) test was employed as a post hoc analysis. Subsequent analysis using Tukey's HSD test unveiled significant

This project is funded by the European Union G.A No 101016906

differences between the gestures and speech groups, as well as between the multimodal and speech groups, both yielding p-values close to 0. Conversely, no significant difference was observed between the gestures and multimodal groups.

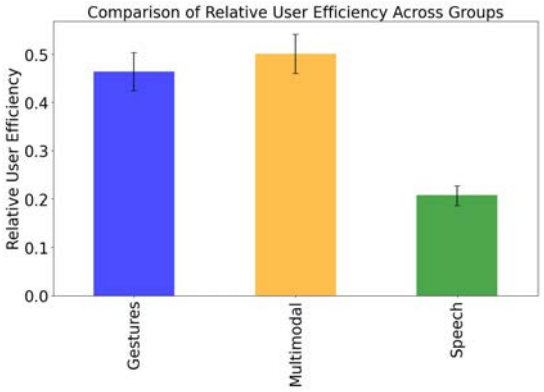


Figure 21. Comparison of relative user efficiency.

Measure	1st Group	2nd Group	Cohen's *d*	Cohen's *U3*
* Rel. user efficiency	Gestures	Speech	1.49	0.93
* Rel. user efficiency	Multimodal	Speech	1.69	0.96
Memorability	Speech	Gestures	1.05	0.85
Memorability	Speech	Multimodal	1.30	0.90
Learnability	Speech	Multimodal	0.66	0.75
* Mental effort	Multimodal	Speech	0.55	0.71
* Physical effort	Gestures	Speech	1.64	0.95
* Physical effort	Multimodal	Speech	1.32	0.91

Table 8. Cohen's and Cohen's U3 of significant difference. Measures marked with an asterisk indicate negative aspects, where lower scores are desirable.

To evaluate the practical significance of these discrepancies, Cohen's ***d*** and Cohen's ***U3*** were computed. These differences can be seen in Table 8. The table can be interpreted as follows: the differences between the 1st group and the 2nd group yield a difference, the effect size of this difference is expressed in Cohen's ***d***, in which a value greater than |0.8| indicates a large effect size, a value greater than |0.5| indicates a medium effect size, and a value larger than |0.2| indicates a small effect size. Furthermore, Cohen's ***U3*** indicates the ratio of participants from the first group that scores higher than the mean of the second group. For example, the comparison between multimodal and speech modalities resulted in a Cohen's ***d*** of 1.69, indicating a large effect size. Additionally, a Cohen's ***U3*** value of 0.96 was obtained, signifying that 96% of the multimodal group will be positioned above the mean of the gestures group.

Perceived competence: The ANOVA test on perceived competence did not achieve statistical significance, yielding a $p - value = 0.34$, thereby precluding further analysis.

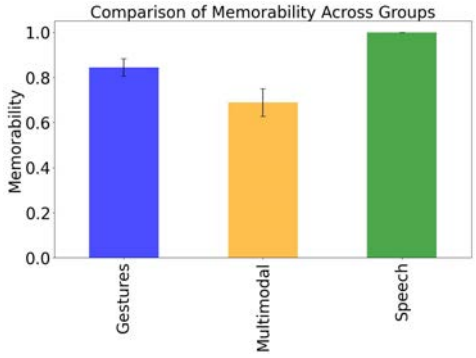


Figure 22. Comparison of memorability.

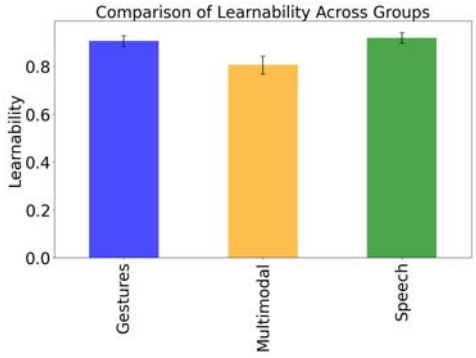


Figure 23. Comparison of learnability.

The Kruskal-Wallis test revealed statistically significant differences between groups for certain variables at $\alpha = 0.05$. The following section details these variables and their results, with practical significance shown in Table 5.

Memorability: The differences in memorability between modalities (see Figure 22) were found to be statistically significant, with a $p-value$ of $7.85*10^{-6}$, which is below the established $\alpha = 0.05$. When significant results are found on a Kruskal-Wallis test, a Dunn's test can be used post hoc to report results among pairwise comparisons. The post hoc analysis yielded a ***p – value of*** 0.01 for the differences between speech and gestures, and a ***p – value of*** approximately 0 for the differences between multimodal and speech.

Learnability: The Kruskal-Wallis test for the differences between means in learnability (see Figure 23) yielded a ***p – value of*** 0.04, which is below the established alpha and thus meets the criteria for statistical significance. In a post hoc analysis, a $p-value$ below the established α was obtained for the differences between multimodal and speech ($p - value = 0.04$).

Mental effort: The differences between mental effort (see Figure 24) expended per modality yielded a ***p –value =*** 0.01. A post hoc analysis showed that these differences were only significant between multimodal and speech, with a ***p – value =*** 0.01.

This project is funded by the European Union G.A No 101016906

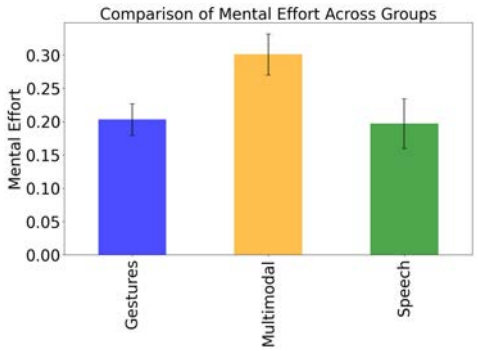


Figure 24. Comparison of mental effort expended on task.

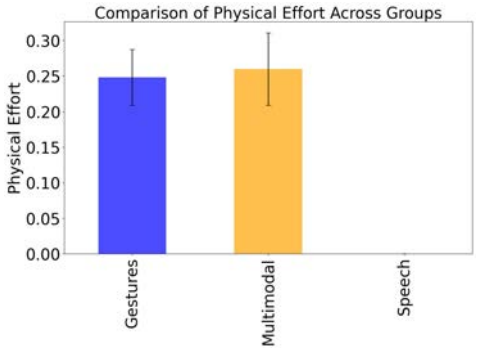


Figure 25. Comparison of physical effort expended on task.

Physical effort: The differences in physical effort (see Figure 25) expanded on the task yielded a $p-value = 2.41*10^{-8}$ and were thus statistically significant. After applying Dunn's test, the differences were only found to be significant between multimodal and speech (***p –value of*** roughly 0) and between gestures and speech ($p-value$ of roughly 0).

4.1.9 Conclusions

Creating a multimodal communication framework aims to enhance human-robot interaction in scenarios where unimodal communication may fall short. However, user analysis indicates a preference for unimodal interactions, specifically speech, regarding usability.

Although the single measure of usability yielded no statistically significant differences between groups, several facets of usability did. Speech-only interactions were favoured in several subcomponents: relative user efficiency, memorability, learnability, mental effort, and physical effort. These results align with previous research [5]. Contrary to expectations, there were no significant differences in trust, perceived competence, or perceived discomfort.

Regardless of user opinion, there are situations where employing the speech-only modality is simply impossible. In these situations, users suggested improvements to the gesture recognition system in terms of (i) choice of gestures, (ii) smoother transition from one gesture to another, (iii) distinction between gestures for long-lasting actions (e.g., move on indefinitely) and fine-grained ones (e.g., move on just a bit), and (iv) reducing the number of needed gestures (e.g., implementing an efficient collision avoidance mechanism).

This project is funded by the European Union G.A No 101016906

Future improvements to the system may alter user opinions. For instance, speech commands experienced a latency of approximately 55 milliseconds, which users noted and suggested should be reduced for a smoother experience. These issues underscore the need to consider environmental and participant-specific factors in HRI experiment design and interpretation.

Finally, participants expressed diverse preferences for speech-only, gestures-only, multimodal interactions, or even using a controller. Future studies could experiment with a within-subjects design, in which users participate in all three modalities, to analyse real preferences after users experience all modalities.

4.2 Human-Robot Awareness

"Human-Robot awareness is based on a high-level specification of the environment and tasks, for example by using Petri Nets describing agronomic processes. The suitability of process models to describe actual reality can be captured by using Conformance Checking, which is a common task in Business Process Management (BPM) providing a quantitative measure of how well a process model adheres to actual behavior. This measure is usually called fitness. The higher is the fitness value (comprised between 0.0 - very poor fitness - and 1.0 – perfect fitness) This mechanism can be used at runtime to detect deviations but can be used also to evaluate the human awareness mechanisms. Evaluation will be then performed by asking humans to carry on complex activities together with robots, and by measuring how well they can be described by process models defined in the context of the project. For our project we require a fitness of at least 0.7 on all the high-level tasks."

In this section, we evaluate the performance of the methodology proposed in D6.3. With respect to that methodology, we separately assess the process extraction and process awareness phases. In addition to the indicated fitness benchmark, which relates only to process extraction, we also measured how well the proposed approach is able to answer questions about the status of the activity. This second evaluation was initially not planned but it adds an additional perspective to evaluate the quality of the proposed approach.

4.2.1 Evaluating Process Extraction

To convert user commands from natural language into LTL formulas, we utilized Lang2LTL. We selected a set of 23 symbols: 4 for locations, 13 for conditions, and 6 for actions. Each symbol is implemented as a specific fluent within the PDDL domain, allowing us to ground the formulas in the planning context. For each symbol, we chose a brief natural language phrase, which is included in the landmark file used by the tool to associate the symbols with referring expressions during grounding.

We created a dataset comprising over 500 natural language sentences relevant to our domain. Approximately 200 sentences express navigation patterns (such as Visit, Sequenced Visit, Ordered Visit, Strict Ordered Visit, and Global Avoidance) using only location symbols. The remaining 300 sentences refer to generic tasks using all symbols in P, corresponding to patterns like Bound Delay, Delayed Reaction, Prompt Reaction, Wait, Past Avoidance, and Future Avoidance.

To expedite the dataset creation process and minimize bias, we enlisted the help of GPT-3.5. For each pattern, we provided GPT with a description of the pattern, an overview of the symbols available for use, and three example pairs (sentence, LTL formula) to guide the generation of additional data. The generated data was then examined to ensure quality and relevance.

4.2.2 Evaluating Process Awareness

Quantifying the degree of awareness in a system is a complex challenge. How can we determine if our system is aware ***enough***? How can we compare the awareness levels of different systems? These questions must be addressed when designing an evaluation mechanism.

The objective of our work is to assess whether an LLM-driven embodied agent, operating in a collaborative scenario with humans and robots, can understand and provide explanations regarding its policy and the actions it has performed, is performing, or will perform. To evaluate this, we conducted a statistical analysis by computing the percentage of correctly predicted fluents across various states.

We focus our study on three primary evaluation scenarios: assessing awareness of what the system is ***currently*** doing, what it has done ***in the past up to the current state***, and what it will do ***in the future starting from this state***. For each scenario, we defined six questions to input into the LLM (along with the policy and PDDL files, as described in detail in D6.3) to elicit responses from the system. During the evaluation, a sentence is uniformly sampled from the corresponding set based on the scenario being considered. This approach minimizes bias from specific input patterns that may influence the responses.

Evaluating Current State: To compute the percentage of correctly predicted fluents against the actual ones in a given state, we need a method to convert the natural language answer provided by the LLM into a set of fluents. For this purpose, we employed another instance of GPT-4, prompting it to function as a $\mathit{FluentExtractor}(\cdot)$. This function takes the following inputs: ***(i)*** a natural language sentence that describes a state (e.g., ***I am currently in the initial location***); ***(ii)*** a set of admissible (ungrounded) fluents F (i.e., all the fluents present in the PDDL files in general form, e.g., ***(robot-at ?var1)***); ***(iii)*** a set of admissible objects (i.e., all the objects present in the problem file, e.g., ***loc0***); and outputs a set of (grounded) fluents $\hat{f}$ derived from the state description (e.g., ***(robot-at loc0)***). Let f represent the actual set of fluents describing a PDDL state, obtained by applying the ADD and DEL sets[1] of all actions taken to reach the current state. By our definition of $\hat{f}$ derived from the $\mathit{FluentExtractor}$, it always holds that $\hat{f} \subseteq f$. Based on these premises, we define our accuracy metric Γ as follows:

$$\Gamma = \frac{|\hat{f} \cap f|}{|f|}$$

Evaluating Past and Future States: To assess awareness regarding what has been accomplished and what will occur in the future, we cannot use the metric Γ in its current form. The robot's response may include concepts that refer to fluents that, in the case of the past, were once achieved but

[1] For an action in a planning domain, the ADD and DEL sets represent the predicates that are added (i.e., those that become true) and those that are removed (i.e., those that become false) in the status of the world, respectively.

subsequently deleted by later actions, or in the case of the future, have not yet been achieved. Therefore, applying Γ directly would yield an unfair result that underestimates the actual awareness. To address this, we modify the formulation of Γ to account for these factors. We average the metric over all states that have occurred from the initial state up to the current one (for the past) or over all states that will occur from the current state up to the goal state (for the future). Formally, we define:

$$\Gamma_{past} = \frac{1}{t}\sum_{i=1}^{t}\frac{|\hat{f} \cap f_i|}{|f_i|} \quad 1 < t \leq T$$

$$\Gamma_{future} = \frac{1}{T-t}\sum_{i=t}^{T}\frac{|\hat{f} \cap f_i|}{|f_i|} \quad 1 \leq t < T$$

where t is the timestamp of the current state, and T is the timestamp of the goal state. It is important to note that the planner operates offline, meaning the policy is computed before execution begins. Thus, assuming knowledge of the policy in advance is valid, particularly in the case of evaluating future states. Additionally, to compute Γ, we must ***determinize*** the policy by removing all ***non-deterministic*** decision nodes, grounding them with one of the possible events that can occur.

4.2.3 Experiments

Process Extraction: We assess the fitness of process extracted from Lang2LTL on sentences generated for the CANOPIES domain through three distinct experiments.

1) Vanilla Lang2LTL: In this test, we utilize Lang2LTL with: ***(i)*** the pre-trained symbolic model available online, ***(ii)*** the standard Referring Expressions Recognition (RER) prompt that includes 16 examples, and ***(iii)*** our specific landmark file tailored to the domain.
2) Lang2LTL with augmented RER prompt (11 symbols): For this experiment, we again employ Lang2LTL using: ***(i)*** the pre-trained symbolic model, ***(ii)*** an enhanced RER prompt containing 34 examples, which includes 18 instances of RER in non-navigational generic task sentences, and ***(iii)*** our landmark file. Notably, this RER prompt uses sentences not in the test set and incorporates only 11 of the 23 symbols in our domain (comprising 4 location symbols, 4 condition symbols, and 3 action symbols).
3) Lang2LTL with augmented RER prompt (18 symbols): This experiment closely mirrors the previous one, but we further support Lang2LTL in the RER phase with an enhanced prompt that still consists of 18 sentences. This time, however, it encompasses nearly all symbols in our domain—18 out of 23 (the 4 location symbols, 8 condition symbols, and all 6 action symbols).

Experiment	Navigation tasks	Generic tasks
Vanilla Lang2LTL	86%	22%
RER-prompt 1	**86%**	77%
RER prompt 2	85%	**88%**

Table 9. Fitness of LANG2LTL on navigation and generic tasks.

The results are presented in Table 9. As shown, Vanilla Lang2LTL exhibits poor performance on generic tasks. This shortfall is primarily due to the system's difficulty in recognizing expressions in sentences that refer to actions or conditions. In contrast, enhancing the RER prompt leads to a notable improvement in performance for generic tasks, while navigation tasks remain stable. The assistance provided to the tool in identifying referring expressions correlates positively with

performance. The best outcomes arise from using the RER prompt with examples covering 18 out of 23 symbols, surpassing the highest results achieved in navigation tasks.

Upon reviewing the incorrect predictions from the last experiment, we discovered that, aside from a few grounding errors, many formulas were incorrect even when the RER and grounding were accurate. Thus, achieving 88% fitness appears to be the upper limit for generic tasks without retraining the symbolic translation module, relying solely on optimizing the landmark file and RER prompt. Nonetheless, the symbolic module has demonstrated strong robustness, effectively handling non-navigational tasks when adequately supported by the other modules.

Process Awareness: To assess the agent's awareness, we conducted 30 experiments for each of the three cases, totaling 90 experiments. As outlined in Section IV-B, at the beginning of each run, we sample a question from a set of six distinct (but similar) sentences, which will then be input to the LLM. A critical consideration is: ***when*** do we query the LLM? The human user's question could arise at any point during the policy execution—whether immediately after the robot begins deliberating, just before it reaches the goal, or at any moment in between. To simulate this behavior, we decided to query the LLM after each action execution in the determinized plan, with a probability pp that increases after each action. This approach ensures that a question will eventually be posed before the execution concludes (or immediately upon completion at the latest). We initialized $p_0 = 0.25$, and updated p_t after every action as follows:

$$p_t \leftarrow \min\left(1, \frac{1 - p_0}{length(plan)} + p_{t-1}\right)$$

Category	Γ
What are you doing now?	0.77±0.22
What did you do so far?	0.85±0.16
What are you going to do next?	0.58±0.21

Table 10. Evaluations of the three considered settings.

The only other hyperparameter in this context is the ***temperature*** for both the LLM and the $FluentExtractor$, which we set to $1e-7$ to ensure deterministic output.

The results of the experiments are summarized in Table 10. The high performance observed in the current and past scenarios is unsurprising; knowing all actions executed thus far, along with the complete PDDL specification, enables the agent to effectively explain the rationale behind its actions and its current state. The evaluation of the future scenario also yielded expected results: while the robot executes actions, it cannot predict future outcomes due to environmental nondeterminism. Nevertheless, it manages to estimate, on average, approximately 60% of the fluents that will define the world at some point in the future.

Evaluation with human users: In addition to a quantitative and automatized evaluation, we have also evaluated the system with human users in the virtual environment as done with human robot communication modalities (see Section 4.1). In particular, we have asked 5 researchers to experiment with the system proposing a high-level activity in natural language and then asking questions about

This project is funded by the European Union G.A No 101016906

the progress of the activity, while following the robot during the activity. The researchers were to provide qualitative feedbacks about the proposed approach. The feedbacks were combined to define possible future implementations:

- Despite the positive results, the types of queries posed by humans may demand reasoning capabilities that exceed the current limitations of LLMs. A suggestion here was to investigate the employment of LLMs as agents capable of using tools. This would enable LLMs to tackle complex problems and reduce hallucinations by generating action plans that include the invocation of external services (tools). Such tools can access external information (e.g., documents, calendars) or interact with the virtual or physical world (e.g., robotic arms).
- Regarding process extraction, the component currently employed is limited to specific patterns of LTL formulas, which particularly fits the agricultural scenario, where activities can be loosely specified. In other situations, more structured activities must be defined, which can be more easily described using other formalisms.

This project is funded by the European Union G.A No 101016906

5 Conclusions

This deliverable reported the outcome of the validation of the human-robot interaction functionalities during harvesting and pruning as per Tasks 9.2, 9.3 and 9.4. Each section reported for each of the KPI defined in deliverable D2.3 the corresponding experimental evaluation in the field or in the virtual environment. Notably, all the proposed KPIs have been satisfied and, in most cases, outperformed.

This project is funded by the European Union G.A No 101016906

6 List of attached videos

In the following a list of videos, which demonstrate the tasks 9.2, 9.3 and 9.4 alone and integrated:

- In the video available at the address https://uniroma3.sharepoint.com/:v:/s/project-canopies/EVw1Yn5d7fNEjfONmAHnXNMBrxlO-Ou6enkkgcqGo2Karg?e=oPRGx8, the human robot communication, perception and manipulation functionalities are combined to have a conversation with the human user finalized to harvesting. At https://uniroma3.sharepoint.com/:v:/s/project-canopies/EW4hSsamjilNj6wl_MY5D7cBXwUBUM_V-sQJCntHpbCYVw?e=pVIPOA instead, a shortened version focusing only on harvesting driven by human robot communication.
- In the video available at the address https://uniroma3.sharepoint.com/:v:/s/project-canopies/EZj5RpJa1SlDjlDWoJCxoIMBQdBdQEfiIQ3sJmNRR8iZ7w?e=ivmNxd, the human robot communication in conjunction with navigation functionalities is shown.
- In the video available at the address https://uniroma3.sharepoint.com/:v:/s/project-canopies/EYCx_oDJ7SRLqsOow0MOLukBGIfSHZ8UZM1SO8gSeSi-4g?e=hSvvip, the Human-Robot Awareness functionality is shown in isolation, detached from the Human-Robot communication module.
- In the video available at https://uniroma3.sharepoint.com/:v:/s/project-canopies/EZY1EJ3CAmZNuKVnDVp9K6kBy2kjDqQ653s0gPxpr-tTlg?e=sfC6DY, the basic harvesting operation is shown, whereas https://uniroma3.sharepoint.com/:v:/s/project-canopies/EdesqeC9OiZGpmgGLS7wwaMBT_eBGjTl0cNKjo8Mw3WkQA?e=ShltBK contains a zoom in on the manipulator.
- In the video available here https://uniroma3.sharepoint.com/:v:/s/project-canopies/EZ0ukGhXDQpKv1OWQJiMifkBLQ89_cAeWG4rusYbuODMeg?e=1lmn8H, the focus is on showing safe operations in the presence of humans.
- In the video available at https://uniroma3.sharepoint.com/:v:/s/project-canopies/ERA-YeJSM3lBoVk1swIOc48B6jFePJWE9rEFvyqTygDN6w?e=2486UA, demonstration of the state machine for the harvesting operation, also with the assistance of a human, is shown. Here the same is shown in combination with human-robot interaction https://uniroma3.sharepoint.com/:v:/s/project-canopies/Ea55OvAiqOVOu893qfIqwiwBqTBao4s4umRvsoyX5-WKcw?e=UAipI9.
- In the videos available at https://uniroma3.sharepoint.com/:v:/s/project-canopies/Eb_qEwIJ1JNNvoD_jA2OJHMBNG44phCVFG-61C7wivBW1w?e=kv8u3n and https://uniroma3.sharepoint.com/:v:/s/project-canopies/Ef5gn2wbDPlJj2mQAE0ulXcBs4xA_76omvoZTW0zwq2qdw?e=BIUxDm, the pruning activity is shown in with or without combination with implicit human-robot communication.

7 References

[1] Botterill, Tom, Scott Paulin, Richard D. Green, Samuel Williams, Jessica Lin, Valerie Saxton, S. Mills, Xiaoqi Chen and Sam Corbett-Davies. "A Robot System for Pruning Grape Vines." Journal of Field Robotics 34 (2017).

[2] Abhisesh Silwal, Francisco Yandun, Anjana Nellithimaru, Terry Bates, and George Kantor. Bumblebee: A path towards fully autonomous robotic vine pruning. arXiv preprint arXiv:2112.00291, 2021.

[3] Qinghua Yang, Yuhao Yuan, Yiqin Chen, and Yi Xun. Method for detecting 2d grapevine winter pruning location based on thinning algorithm and lightweight convolutional neural network. International Journal of Agricultural and Biological Engineering, 15(3):177–183, 2022.

[4] Theophile Gentilhomme, Michael Villamizar, Jerome Corre, and Jean-Marc Odobez. Towards smart pruning: Vinet, a deep-learning approach for grapevine structure estimation. Computers and Electronics in Agriculture, 207:107736, 2023.

[5] Julian Abich and Daniel J. Barber. 2017. The impact of human-robot multimodal communication on mental workload, usability preference and expectations of robot behavior. Journal of Multimodal User Interfaces 11 (2017), 211–225. Issue 2. https://doi.org/10.1007/s12193-016-0237-4

[6] G. Coll-Ribes, I. J. Torres-Rodriguez, Antoni Grau, Edmundo Guerra, Alberto Sanfeliu, "Accurate detection and depth estimation of table grapes and peduncles for robot harvesting, combining monocular depth estimation and CNN methods". Computer and Electronics in Agriculture. Vol. 215, December 2023, 108362. https://doi.org/10.1016/j.compag.2023.108362

[7] Alberto Sanfeliu. 2024. Data for publication. CANOPIES Grape bunch and peduncle dataset. https://doi.org/10.5281/zenodo.14234044

This project is funded by the European Union
G.A No 101016906

8 Appendix: Harvesting benchmark

The following images show several cases regarding the harvesting capabilities of the farming robot reporting successful and failing harvesting procedures. Each couple of images shows the robot in pre-harvesting and post-harvesting phases.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase. On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase. On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left , the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left , the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING

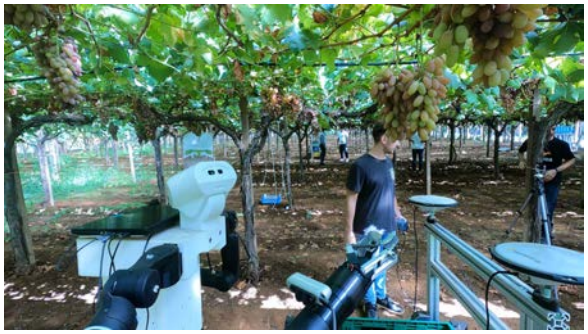


On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union
G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

SUCCESSFUL HARVESTING





This project is funded by the European Union
G.A No 101016906

SUCCESSFUL HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.

This project is funded by the European Union
G.A No 101016906

FAILING HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.
The peduncle is not between the fingers of the gripper.

FAILING HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.
The peduncle is not between the fingers of the gripper.

This project is funded by the European Union
G.A No 101016906

FAILING HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.
The peduncle is not between the fingers of the gripper.

FAILING HARVESTING



On the left, the pre-grasping phase.



On the right, the grasping phase.
The gripper hits the berries and does not reach the peduncl