PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Giuseppe Rauso

Cycle: XXXIX

Training and Research Activities Report

Year: Second

Tutor: prof. Albero Finzi

Co-Tutor: prof. Vincenzo Lippiello

Date: October 29, 2025

Training and Research Activities Report

PhD in Information Technology and Electrical Engineering

Cycle: XXXIX **Author: Giuseppe Rauso**

1. Information:

- **PhD student: Giuseppe Rauso**
- **DR number: DR997198**
- **Date of birth: 30/09/1998**
- **Master Science degree: Computer Science University: University of Naples Federico II**
- **Doctoral Cycle: XXXIX**
- **Scholarship type: PNRR Partenariato Esteso PE15**
- **Tutor: Alberto Finzi**
- **Co-tutor: Vincenzo Lippiello**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
How To Boost Your PhD	Course	18	5	08/01/2025 - 12/02/2025	Prof. Antigone Marino	Y
Can we Rely on AI? Reliability Issues in Artificial Neural Networks and Potential Solutions for Autonomous Vehicles	Seminar	1	0.2	16/01/2025	Dr. Edoardo Giusto	Y
Porte Aperte – Open Day	Tutorship	4	0.16	11/02/2025	Prof. Alfonso William Mauro	Y
Robot Autonomy among Decision-Making Agents	Seminar	1	0.2	15/04/2025	Prof. Fabio Ruggiero	Y
Planning and Navigation	Course	48	6	03/2025 - 06/2025	Prof. Riccardo Caccavale	Y
Robotic Manipulation @ Vanvitelli Robotics Lab: A bird's eye view on the last 5 years	Seminar	2	0.4	18/06/2025	Prof. Mario Selvaggio	Y

Superconducting Radio Frequency Cavities for Quantum Computing and Communication	Seminar	1	0.2	24/06/2025	Prof. Edoardo Giusto	Y
Trusted Execution Environments for QPUs	Seminar	1	0.2	27/06/2025	Prof. Edoardo Giusto	Y
Radar Cross-Section Estimation and Measurements	Seminar	4	0.8	17/10/2025	Prof. Amedeo Capozzoli, Prof. Claudio Curcio, Prof. Angelo Liseno	Y
Guardians or Threats? AI at the Frontlines of Cybersecurity	Seminar	4	0.8	17/10/2025	Prof. Antonia Maria Tulino	Y
AI Powered User interface design	Seminar	4	0.8	24/10/2025	Prof. Antonia Maria Tulino	Y
Quality of Services	Seminar	4	0.8	28/10/2025	Prof. Antonia Maria Tulino	Y

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**0**	**7**	**0**	**7**
Bimonth 2	**5**	**0.2**	**7**	**0.16**	**12.36**
Bimonth 3	**0**	**1**	**7**	**0**	**8**
Bimonth 4	**6**	**1**	**6**	**0**	**13**
Bimonth 5	**0**	**0**	**6**	**0**	**6**
Bimonth 6	**0**	**3.5**	**11**	**0**	**14.5**
Total	**11 (34 total)**	**5.7 (12.3 total)**	**44 (79 total)**	**0.16 (0.16 total)**	**60.86 (125.46 total)**
Expected	**10 – 20 (30-70 total)**	**5 – 10 (10-30 total)**	**30 - 45 (80-140 total)**	**0 – 1.6 (0-4.8 total)**	

3. Research activity:

During the second year of my PhD, I continued the research directions I began exploring in the previous year, focusing on incremental learning by imitation for robotic manipulation tasks and on multimodal attentional mechanisms in reinforcement learning. The core idea behind the research carried out revolves around incremental and modular learning from demonstration and the study of multimodal attentional mechanisms in robotics. In particular, the research focused on application domains such as robotic manipulation, exploration, and navigation. Learning from demonstration is especially useful for accelerating the acquisition of complex skills by leveraging the knowledge of an expert. In the context of manipulation, this is particularly valuable, since with the appropriate instrumentation and interface, it is possible to demonstrate manipulation tasks that are easy for a human to perform but difficult for a robot to learn without guidance. When combined with a modular and incremental methodology, this approach enables the learning of structured tasks by focusing on subtasks and providing specialized demonstrations, while also allowing the composition of primitive skills to formulate new tasks. Attentional mechanisms, on the other hand, have been applied across many different areas of machine learning, and their use has become increasingly widespread thanks to the success of the Transformer architecture—initially developed in the field of Natural Language Processing (NLP) but now extended to a wide range of domains with the advent of Foundation Models. These approaches are highly effective in improving learning efficiency and enhancing focus on the most relevant features of the observed data. The main research focus over these two years of the Ph.D. has been the application of such mechanisms in the context of language-conditioned reinforcement learning and the learning of exploration and object retrieval policies, while simultaneously grounding linguistic elements in the environment. This not only improves the agent's performance but also provides a degree of interpretability of the learned policy.

3.1. Integrating Demonstration-Based Incremental Learning with Symbolic Planning

Regarding the first research area, I extended the ideas developed in my master's thesis and pursued during the first year, centered on incremental learning from demonstrations. In particular, I proposed a framework that, in addition to incorporating this methodology through simulated environments

developed in Unity, integrates symbolic planning based on Hierarchical Task Networks (HTN) and a behavior repository that stores both learned and predefined skills, each associated with a symbolic representation. This repository enables the combination of skills to perform long-horizon robotic tasks. The framework is designed to be modular and incremental, and it is compatible with the approach presented last year on learning robotic manipulation tasks from demonstrations in virtual reality using an anthropomorphic end-effector. Indeed, the policies learned in this way can be stored in the repository and later used as operators for task composition.
The learning component is structured to be as robot-agnostic as possible. For instance, in training grasp-and-lift and place tasks, I employed general observations that do not depend on the specific gripper used, and the robot arm itself was not considered during training. This allowed us to switch to a different simulator during the evaluation phase—CoppeliaSim—closer to real-robot integration, and to effectively use a different robotic arm and gripper. Likewise, object observations are kept general, relying only on their bounding boxes. Naturally, manipulation can be made more complex by introducing end-effectors equipped with tactile sensors, thus balancing the generality of bounding-box observations (as discussed in last year's work). This can be achieved simply by adding new skills—such as learned policies—to the behavior repository, so that they can be invoked by the planner depending on the current situation.
This approach provides a foundation for incrementally adding new skills and enhancing the robot's capabilities while maintaining modularity. It therefore allows new functionalities to be introduced at different levels of granularity. This work was presented at the AIRO 2025 workshop, held during the I-RIM 3D 2025 conference, and has also been submitted to the *Robotics and Autonomous Systems* journal, where it received a minor revision in the first review round and is currently under the second round of review.

3.2. Multimodal and Task-Aware Attention in Robotic Learning

The second line of research concerns the study of attentional mechanisms for exploration and retrieval tasks in robotics. I continued to refine the idea I introduced in my first-year work, which focused on a methodology for computing text-visual attention and task attention in the learning of "go to" and "pick up" tasks within BabyAI environments. In this context, tasks are specified in natural language following a well-defined syntax and structure.
The research conducted this year focused both on improving the architecture of the proposed models—now more robust and with enhanced grounding capabilities—and on transferring the learned policies to real robots. The improvements achieved in both the architecture and the methodology—covering single-task and multi-task settings with task attention—were included in a paper submitted to the AAAI conference. Although the paper was not accepted, mainly due to the lack of comparisons with more recent Transformer-based models (whose hardware requirements often make such comparisons impractical), I collected the reviewers' feedback and plan to prepare an extended journal version of the work over the next year, to be submitted to IEEE Transactions on Cognitive and Developmental Systems.
Another part of the research activity focused on transferring the policies learned in the BabyAI grid-based simulation to a real rover. Together with colleagues from PRISMA Lab and other universities, I conducted experiments aimed at creating a more abstract grid-based representation of the environment perceived by the rover, making it compatible with the observations used in BabyAI. This enabled the use of the learned policies, leveraging the proposed attentional mechanisms, to perform reaching and

retrieval tasks. The experiments yielded successful results, and a journal article is currently in preparation for submission in the coming year.

3.3. Single- and Multi-Robot Space Exploration and Foundation Models in Robotics

As part of the SpaceItUp project, I carried out research activities in the context of both single-robot and multi-robot space exploration. Together with other colleagues from the PRISMA Lab, I explored methodologies for collaborative lunar exploration with heterogeneous robots, and the results were included in the paper accepted at the CEAS-AIDAA conference. In addition, I collaborated in supervising students during their thesis work, which focused on single-robot lunar exploration for sample retrieval and return-to-base tasks using reinforcement learning.
Within the same project, I contributed to the preparation of Deliverable D8.2.2, including research contributions related to multimodal attentional mechanisms for robotic exploration tasks and the incremental learning-by-demonstration approach for manipulation tasks.
In addition, this year I deepened my study of literature on the use of Large Language Models (LLMs), Vision-Language Models (VLMs), and Vision-Language-Action Models (VLAs) in robotics. Research in this area is highly active, and the capabilities of such models often enable zero-shot adaptation to new situations, thanks to the knowledge they acquire during training. However, fully end-to-end approaches still appear suboptimal due to the so-called “hallucination” problem. Many recent works therefore focus on combining multiple modules based on these models to accomplish more complex tasks, which also allows for greater control over potential failures.
We addressed these topics in collaboration with colleagues from Universidad Carlos III de Madrid in the paper submitted to the ICRA conference, which explores a method for evaluating the feasibility of tasks expressed in natural language in the context of assistive robotics using Multimodal Large Language Models (MLLMs). In the future, other models trained on robotic domains—such as Gemini Robotics-ER or Vision-Language-Action (VLA) models—will also be explored, representing a promising direction for further research.
Another area I explored is multi-agent reinforcement learning, studying algorithms such as MA-POCA, proposed by Unity and integrated into ML-Agents—a framework I have extensively used in several previous works. The knowledge acquired in this domain will be valuable next year for research related to multi-robot spatial exploration.

3.4. Future Work

As a natural extension of the work carried out this year, a key direction will be to combine the methodologies I have studied and proposed in my research on multimodal attentional mechanisms and the learning of robotic manipulation tasks from demonstrations. In both cases, the proposed approaches aim to make the most effective use of the available data and to focus on the most relevant aspects of the task to be learned, thereby improving training efficiency and yielding more effective policies.
Furthermore, the grounding capabilities enabled by the attentional mechanisms can be leveraged in more complex environments, possibly using realistic visual inputs, to guide the robot toward relevant elements in the scene. These attentional cues could then be integrated with the manipulation skills learned incrementally through demonstrations and reinforcement learning.
To this end, it will be particularly interesting to study and propose methodologies for integrating powerful models such as LLMs, VLMs, and VLAs, which could serve as a bridge between the use of language to define tasks and goals, the perception of the human operator for imitation and monitoring in

collaborative tasks, and the retrieval and execution of learned or predefined behaviors for robotic tasks. These topics will be explored primarily in the context of robotic manipulation and exploration, also investigating adaptation to multi-robot or collaborative scenarios, in line with the SpaceItUp project.

4. Research products:

[1] *Conference paper: G. Rauso, R. Caccavale, A. Finzi, "Combined Text-Visual Attention Models for Robot Task Learning and Execution". In: 23rd International Conference of the Italian Association for Artificial Intelligence. AIxIA 2024, Bolzano. Presented and published.*

[2] *Workshop paper: G. Rauso, R. Caccavale, V. Lippiello, A. Finzi, "Integrating Text-Visual and Task Attention for Language-Guided Robot Learning". In: 11th Italian Workshop on Artificial Intelligence and Robotics co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). AIRO 2024, Bolzano. Presented and published.*

[3] *Journal paper: G. Rauso, R. Caccavale, A. Finzi, "Incremental Learning from Virtual Demonstrations and Task Composition for Robotic Manipulation". In: Robotics and Autonomous Systems (RAS). Submitted, currently in second review round (first round: minor revision).*

[4] *Conference paper: G. Rauso, R. Caccavale, A. Finzi, "Task-Aware Multimodal Attention for Language-Guided Learning and Execution in Robotics". In: 39th Annual AAAI Conference on Artificial Intelligence. AAAI 2026, Singapore. The paper was not accepted and is currently being extended for submission to the IEEE Transactions on Cognitive and Developmental Systems.*

[5] *Conference paper: R. Caccavale and S. Ciaravino, A. Finzi, V. Lippiello, G. Rauso, "A Heterogeneous Multi-robot Framework for Cooperative Lunar Exploration". In: 28th AIDAA International Congress and the 10th CEAS Aerospace Europe Conference. CEAS-AIDAA, Turin. Accepted.*

[6] *Workshop paper: G. Rauso, R. Caccavale, A. Finzi, "A Unified Framework for Incremental Skill Acquisition and Symbolic Task Composition in Robotic Manipulation". In: 12th Italian Workshop on Artificial Intelligence and Robotics co-located with the Italian Institute of Robotics and Intelligent Machines (I-RIM 3D 2025). AIRO 2025, Rome. Accepted and presented.*

[7] *Conference paper: A. Mora, G. Rauso, R. Caccavale, A. Finzi, R. Barber, "Grounded Intent Validation: A Visual-Semantic Framework for Task Feasibility Assessment in Assistive Robotics". In: 2026 IEEE International Conference on Robotics & Automation. ICRA 2026, Vienna. Submitted.*

[8] *Extended abstract: G. Rauso, R. Caccavale, A. Finzi, "Combining Text-Visual and Task Attention for Language-Guided Robot Learning". In: IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots. SIMPAR 2025, Palermo. Poster presentation (not included in the proceedings).*

[9] *Project deliverable*: Contribution to the development of deliverable D8.2.2 – Report on robotic systems and associated technologies & TRL assessment plan for the SpaceItUp project.

[10] *Prototype software*: I extended the simulation environment in Unity to support the recording of demonstrations and the learning of robotic manipulation tasks, including the option to use a gripper without a manipulator to focus exclusively on end-effector motion. Moreover, I developed simulation environments in CoppeliaSim to deploy and evaluate the policies trained in Unity on realistically simulated robots with full joint motion, implementing a sim-to-sim transfer.

5. Conferences and seminars attended

Conference: 23rd International Conference of the Italian Association for Artificial Intelligence. AIxIA 2024, Bolzano, 25-28 November 2024; presented the paper "Combined Text-Visual Attention Models for Robot Task Learning and Execution".

Workshop: 11th Italian Workshop on Artificial Intelligence and Robotics co-located with the the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). AIRO 2024, Bolzano, November 26th 2024; presented the paper "Integrating Text-Visual and Task Attention for Language-Guided Robot Learning".

Workshop: 12th Italian Workshop on Artificial Intelligence and Robotics co-located with the Italian Institute of Robotics and Intelligent Machines (I-RIM 3D 2025). AIRO 2025, Rome, October 17th 2025; presented the paper "A Unified Framework for Incremental Skill Acquisition and Symbolic Task Composition in Robotic Manipulation".

6. Activity abroad:

None

7. Activity in partner companies:

None

8. Tutorship

None