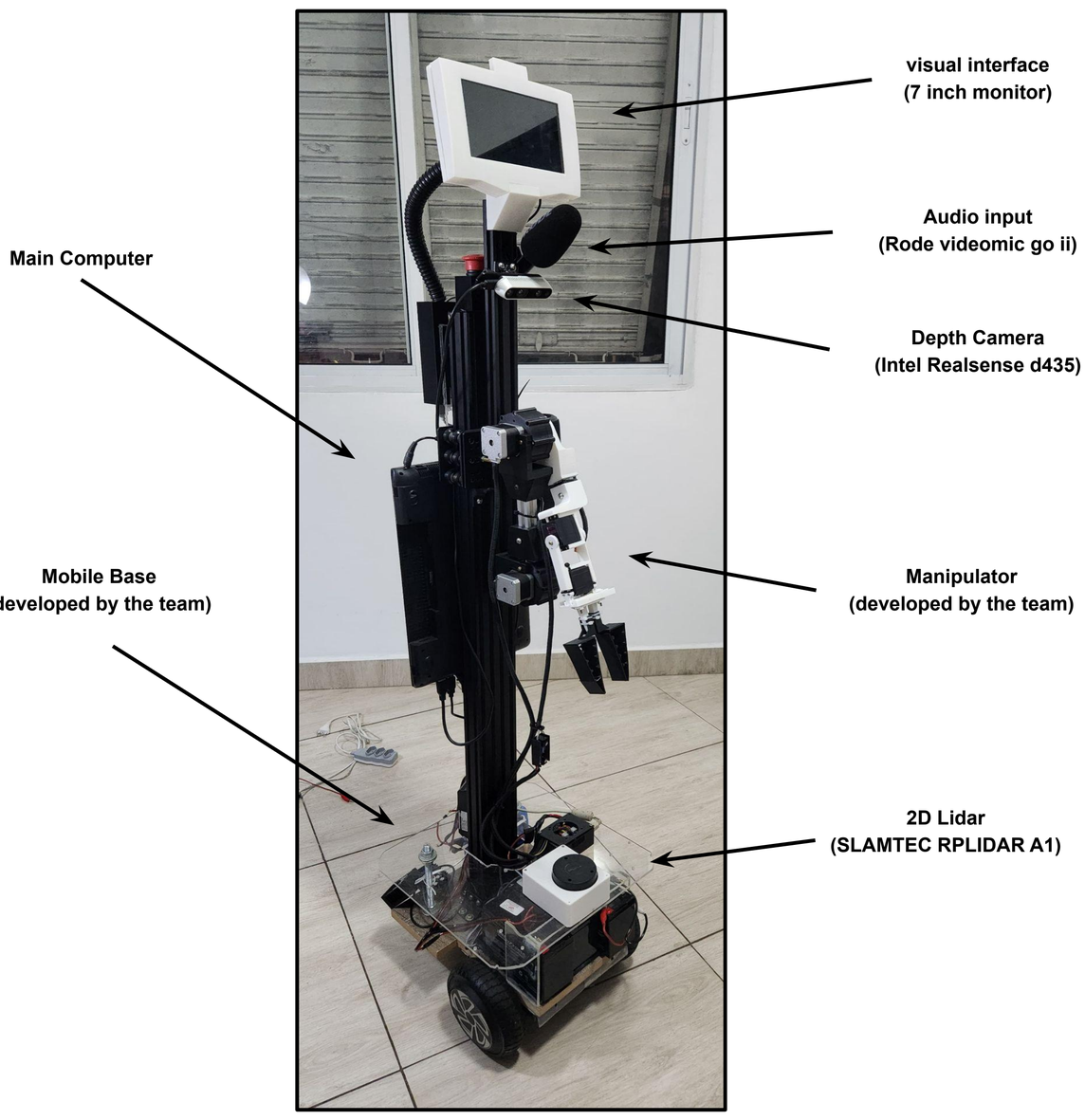
Introducing ATHENA: Autonomous Task Handling and Execution Agent

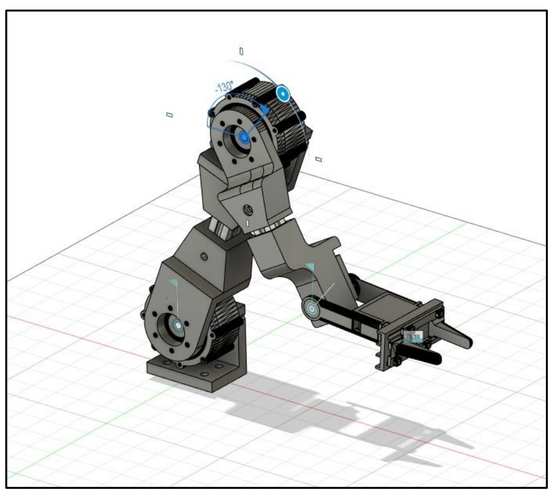
Leonardo C. Neves, Douglas R. Meneghetti,
Rodrigo P. Bechelli, Rodrigo G. G. Piva

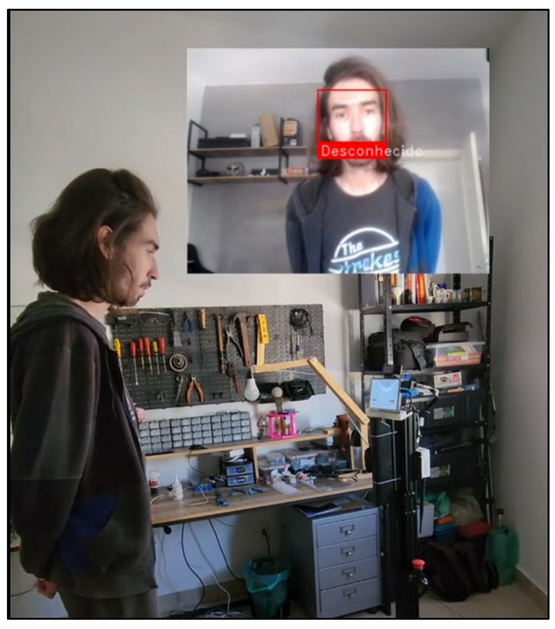
Introduction

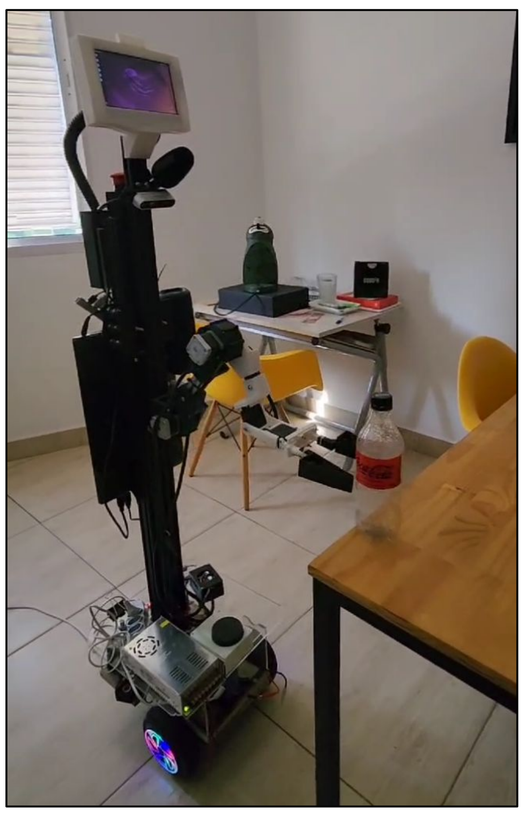
Over the past year, the Ada Robotics team has focused on building a service robot platform equipped with a mobile base, manipulation system, and user interface display. The team established fundamental capabilities including computer vision, autonomous navigation, and speech recognition. This foundation enables the robot to handle both general-purpose tasks and competition challenges.

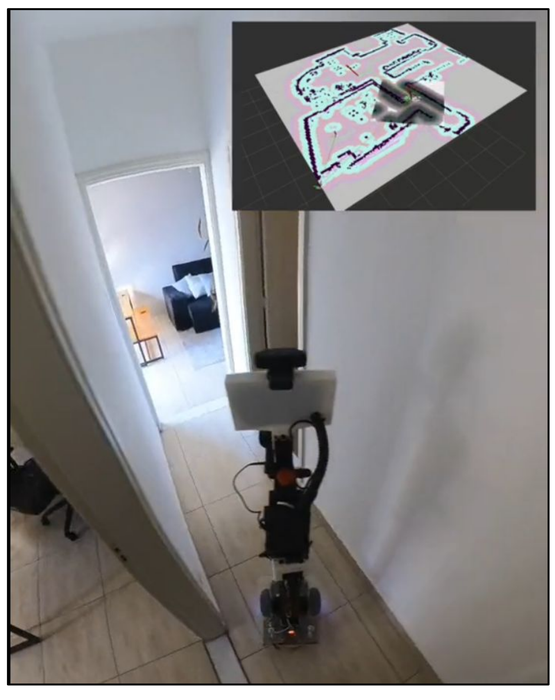
The robot's development serves a dual purpose: participating in service robotics competitions while also opening opportunities for external collaborations and research projects. As a new initiative, the platform represents our first step toward contributing to the advancement of service robotics, with many planned improvements and capabilities yet to be developed.









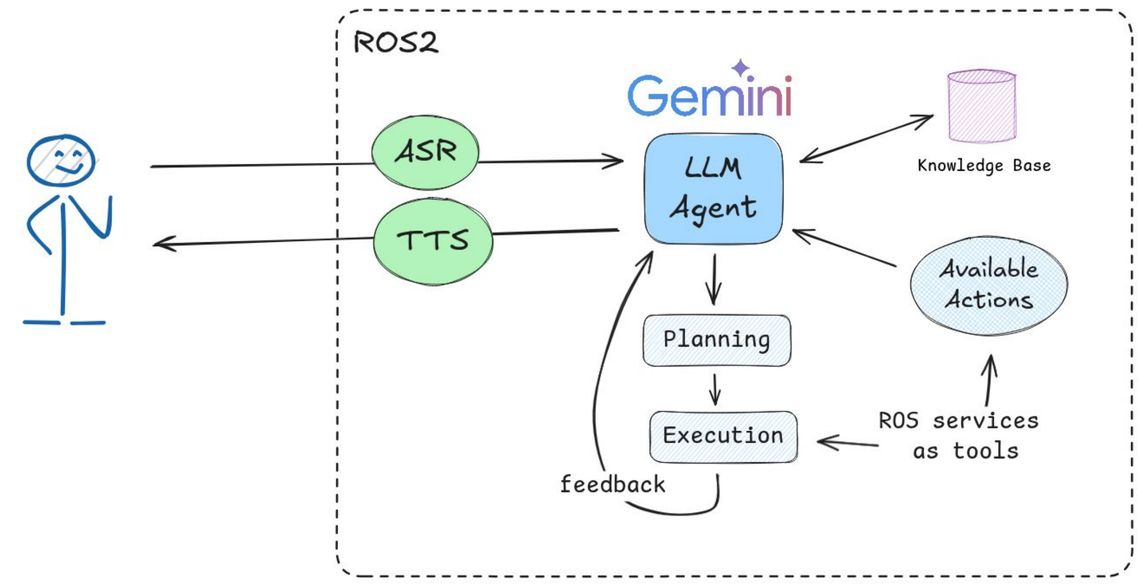


Current Work and Future

Integrating Large Language Models for Decision-Making in Service Robots

This project investigates the integration of Large Language Models (LLMs) to enhance decision-making capabilities in service robots within the ROS2 framework. This system enables robots to interpret user commands and respond intelligently by utilizing Automatic Speech Recognition (ASR) for input processing and Text-to-Speech (TTS) for output, creating a natural and interactive user experience.

At the core of the project, an LLM agent connects to a knowledge base and uses a set of available actions to plan and execute tasks via ROS2 services. The agent processes incoming data and adapts based on feedback, refining its decision-making in dynamic environments. Currently in development, we will released this project as a ROS2 package, offering a modular solution for intelligent task-planning in service robots.



Multimodal System for Service Robot Task Understanding

This system implements a novel approach for service robots to understand and execute natural language commands through multimodal embeddings. The architecture leverages CLIP-inspired shared latent spaces to connect user instructions with robot's visual perception and spatial awareness. The system maintains a vector database of camera frames paired with map positions and dynamic variables. When receiving a command, text input is projected into the shared embedding space to retrieve matching visual frames and metadata. A dual-processing pipeline employing Vision-Language Models (VLM) and Large Language Models (LLM) enables robust task understanding and execution planning, demonstrating effective performance in real-world service tasks while maintaining contextual awareness through its multimodal architecture.