LisTex United
2026 Team Description Paper

Raghav Arora[1], Afonso Certo[2], Rodrigo Coimbra[2], Justin Hart[1], Yoonwoo Kim[1], Pedro U. Lima[2], Duarte Santos[2], Peter Stone[1], and Lingyun Xiao[1]

[1] The University of Texas at Austin, USA
[2] Institute for Systems and Robotics, IST - University of Lisbon, Portugal

Abstract. LisTex United is a new RoboCup@Home team that formed as part of a joint project (FOMO-HODOR) between the Institute for Systems and Robotics at IST, U. Lisbon and UT Austin, with contributions from SocRob@Home and UT Austin Villa @ Home team members. The team focuses on using a humanoid robot (Booster T1 with grippers) to perform the domestic tasks; investigating and addressing the challenges between the current state of the art and widespread deployment humanoid robots in homes. The FOMO-HODOR project will leverage multi-modal Large Language Models (LLMs) in various roles; aiming to harness their full potential within practical general-purpose robotic deployment scenarios, and leveraging their advanced reasoning capabilities for the effective planning of complex tasks. LisTex United will build on the vast experience of its team members in RoboCup@Home to integrate general speech interaction and task planning methods with humanoid-specific topics, such as full-body motion planning, vision-based grasping and locomotion.

1 Introduction

UT Austin Villa @ Home has participated in seven RoboCup@Home competitions. The team's philosophy is to use RoboCup@Home as a springboard for research problems that extend beyond the competition. The systems developed by UT Austin Villa @ Home are also deployed in experiments across UT Austin's Laboratores and in a research program surrounding real-world deployments.

The SocRob@Home team has represented the Institute for Systems and Robotics (ISR-Lisboa) in RoboCup since 1998. Originating from the ISR-Lisboa SocRob (Soccer Robots or Society of Robots) research initiative, the team has participated in multiple RoboCup leagues, including RoboCupSoccer Simulation, 4-Legged and Middle Size, RoboCupRescue Real Robot and, since 2016, RoboCup@Home Open Platform, where it got to the podium twice in 2023 and 2024. SocRob@Home has involved over 100 students, from Bachelor's to Ph.D. levels, post-doc researchers and professors, fostering the integration of task planning, navigation, perception and manipulation towards advanced domestic service robotics.

These two groups have combined their efforts in a research project, featuring a case study on RoboCup@Home to carried out by the LisTex United team. There are significant research challenges that must be met to achieve reliable and contextually aware task planning for robots. Prior attempts at LLM integrations on robots have been relatively simplistic. The project seeks to better integrate the technology into a more cohesive robotics architecture, aiming at a FOundation MOdel for HumanOid DOmestic RObots (FOMO-HODOR).

The impact of the envisioned framework will be substantial. LLMs have not yet been well-integrated into robotics AI architectures, with existing work mostly using them as a substitute for the current state-of-the-art methods. LLMs, however, also lack the strengths of the methods that researchers are attempting to replace with LLMs. LLMs fail on long-horizon planning tasks; can produce inconsistent, unactionable plans; or simply fail to perform well-understood, easily-programmable tasks. The most major promised advantage is an increase flexibility and fault tolerance over classical methods. The framework developed under this research will be transitioned to a variety of robots, performing a variety of tasks, and is likely to have far-reaching impacts for broader and more rapid deployment of robot technologies in real-world scenarios.

2 Software and Scientific Contributions

This section describes the scientific contributions made to enable humanoids to perform Robocup@Home tasks including knowledge representation, semantic perception, whole body manipulation, and locomotion.

2.1 Robot Architecture

The LAAIR robot architecture [1] is designed to support dynamic human-robot interaction in complex environments. The three-layer structure, illustrated in Figure 1, integrates the robot's skill components, such as perception, navigation, and manipulation, with high-level reactive and deliberative control modules. The top layer sequences and executes skills while remaining reactive during execution to maintain responsive human-robot interaction. A central knowledge base facilitates information sharing between system components, providing a unified world model accessible to both reactive and deliberative processes. The deliberative control layer leverages this shared knowledge to reason about the environment and generate plans for tasks that cannot be statically decomposed, following the design described by Jiang *et al.* [1]. This layered architecture provides a flexible foundation that accommodates both cloud-based and onboard planning pipelines, enabling different approaches to be seamlessly integrated within the same overall framework. The team's present research efforts use LAAIR as a starting point to be extended with new components and integrated with foundation-model based approaches. This approach provides us with a reliable and robust baseline that we know to be competition-ready, and well-defined points in the architecture for these integrations.

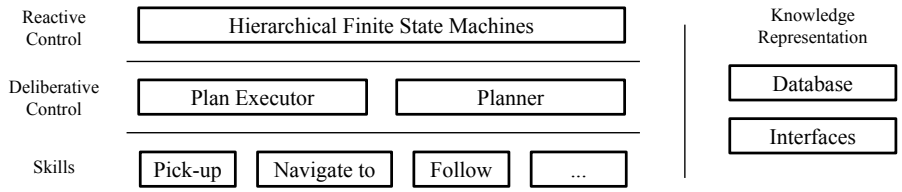


Fig. 1: Implementation of our robot architecture on Booster T1.

2.2 Knowledge Representation and Planning

Our knowledge representation subsystem stores grounded robot knowledge in a SQL database in order to allow for fast access and easy querying. Queries can be formed using custom C++ and Python libraries. For instance, in the *Storing Groceries* and *GPSR* tasks, the knowledge base is used to query object properties such as categories and default locations. The knowledge base can be dynamically updated by our perception system, described below. Fig. 2 shows the knowledge base after the robot has detected a ketchup bottle on the dining table.

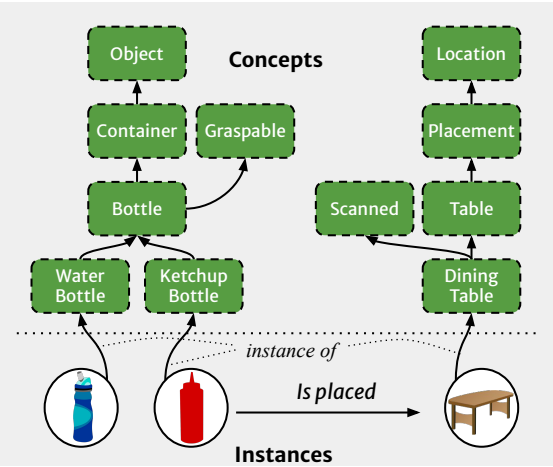


Fig. 2: Visualization of a knowledge base grounded in the robot's perception.

At UT Austin, ongoing work extends this subsystem with an interface based on predicate logic, allowing the knowledge base to be seamlessly integrated into task and motion planning. Core to this approach is the ability to reason about hypothetical objects that are requested by users but not yet perceived by the robot. This capability is crucial to our solution of incomplete commands in earlier versions of the *EGPSR* test. Details of this knowledge representation and

planning system are described by Jiang *et al.* [2]. More recently, Kim *et al.* [3] proposed CoCo-TAMP, which complements the knowledge-based approaches by using LLMs to provide commonsense priors (e.g., likely household locations of requested objects) and guided belief updates to enable partially observable task and motion planning practical in the real world. For the upcoming competition, our system will combine the existing knowledge base with CoCo-TAMP's [3] LLM-derived commonsense priors.

2.3 Human-Robot Communication and Task Planning Using Foundation Models

Reliable human–robot communication is crucial for tasks like *GPSR* and *E-GPSR*, where spoken instructions must be converted from speech-to-text (STT) on the fly and translated into structured commands the controller can execute.

Our current STT pipeline is based on a local version of the Whisper model [4], accessed through the `whisper_ros` ROS 2 wrapper. This implementation supports the use of the Silero voice activity detector [5], which detects the beginning of speech, without predefined start keywords, enabling more natural interactions.

For command understanding we will follow an approach inspired by Liu et al. [6], in which a large language model (e.g., GPT-5) converts the STT transcript into a structured JSON representation of goals for the CoCo-TAMP planner [3] described in subsection 2.2. The LLM is additionally prompted to correct common ASR errors and validate the resulting schema to ensure robustness in online operation.

Speech synthesis is handled by `tts_ros` which interfaces with Coqui TTS or lightweight local engines (e.g. Kokoro). Audio playback through `audio_common` allows the robot to confirm user commands, request clarifications, and provide spoken status updates.

2.4 Manipulation

Segmentation and Grasp Sampling In previous years, the teams relied on YOLOv8 [7] for object segmentation, which required on-site data collection and manual labeling. This year, we instead adopt the foundation vision model Grounded-SAM2 [8], which enables object segmentation directly from natural-language queries, thereby eliminating the need for task-specific data gathering and labeling.

To improve grasp reliability, we replaced AnyGrasp [9] with a diffusion based grasp sampler GraspGen [10] that we found to be better suited to humanoid manipulation. In our pipeline, the segmented target object is combined with the scene point cloud to extract a target-object point cloud, which is then provided to GraspGen to generate a dense set of candidate grasps.

Motion Planning For tasks such as Storing Groceries, the robot must execute a sequence of pick-and-place actions. Because humanoid robots have high-

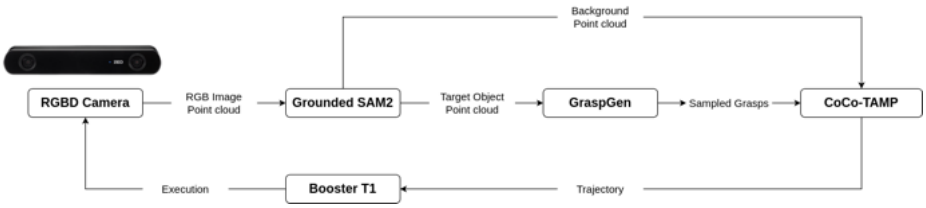


Fig. 3: Overview of manipulation pipeline. RGB-D camera captures an RGB image and point cloud of the scene. Grounded-SAM2 segments the target object from the RGB image and obtains segmented point cloud. The segmented point cloud is passed to GraspGen, which generates a set of candidate grasps. In parallel, the background point cloud is sent to CoCo-TAMP which uses the background point cloud and sampled grasps to plan a feasible manipulation trajectory that can be executed by Booster T1. Execution closes the loop by returning to perception for continued updates and replanning if necessary.

dimensional state spaces and strict balance constraints, motion planning is particularly challenging. To address this, CoCo-TAMP has been upgraded to leverage CuRobo [11], a CUDA-accelerated motion planning library that generates minimum-jerk trajectories. Minimum-jerk motion is important in humanoid manipulation because it produces smooth, continuous trajectories with low acceleration discontinuities, which helps maintain balance.

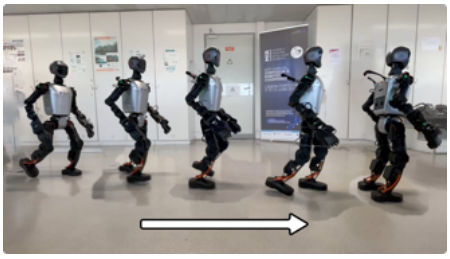
The motion planner first plans as if the humanoid has an omnidirectional base with a lifting column instead of legs. This simplifies the motion planning problem and removes the need for the motion planner to consider the dynamics of the robot and the world. When following the plan, the upper body of the humanoid can directly follow the trajectories from the motion planner. For the lower body, we plan to train a flow matching policy that outputs the trajectory of the lower body conditioned on the head pose and the height of the waist. The role of the trained policy is to balance the robot while following the trajectory from the motion planner.

2.5 Locomotion and Navigation

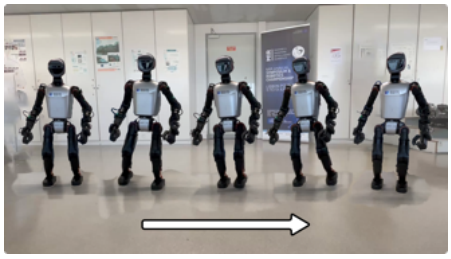
Stable and robust locomotion are critical for the robot to navigate the arena, especially given the added complexity of dynamic bipedal walking on humanoid robots as opposed to statically stable or wheeled systems.

To address this, we chose to develop our own locomotion policy (as showcased in Figure 4), trained using FastTD3 [12]. The qualification video shows the capabilities of the policy deployed on the real robot, including forward, backward, lateral, and rotational motions, as well as a direct comparison with the original Booster T1's motion model, where a noticeable improvement in robustness and gait fluidity is observed.

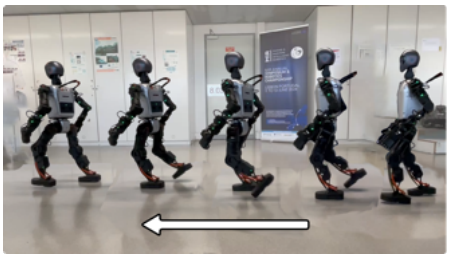
Having achieved stable and robust locomotion, we built a navigation stack on top of it, illustrated in Figure 5, that sends velocity commands to our locomotion



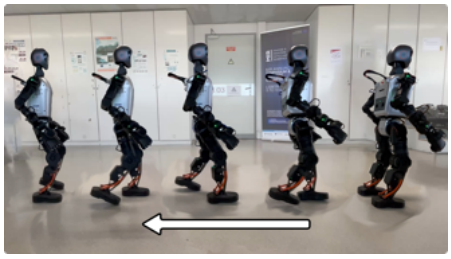
(a) Policy given a forward command.



(b) Policy given a lateral command.



(c) Policy given a forward command.



(d) Policy given a backwards command.

Fig. 4: Developed locomotion policy deployment on the real robot.

policy, enabling the robot to autonomously navigate around the house. Since our system does not rely on wheels for locomotion, wheel odometry is unavailable. Therefore, we use the IMU and RGB-D camera to perform visual odometry. This is achieved through the use of RTAB-Map [13][14], which enables SLAM and localizes the robot within a previously known map. RTAB-Map is capable of detecting loop closures through visual observations, minimizing accumulated mapping errors and allowing the robot to relocalize itself.

Using the generated map and the odometry provided by RTAB-Map, we used the Nav2 package [15] to enable autonomous navigation. As the robot is not equipped with a LiDAR sensor for environment perception, the RGB-D camera is used to generate a point cloud of the surroundings. This point cloud is leveraged to update the costmaps and detect unexpected obstacles in the environment. Since obstacle perception relies on the camera, we adopt a differential drive–based motion model in the controller, restricting the commanded motions to forward translation and in-place rotation, which enables the robot to safely traverse the environment.

3 Conclusions

LisTex United is the RoboCup@Home team resulting from a joint research project between UT Austin and ISR-Lisboa, funded by the UTAustin-Portugal

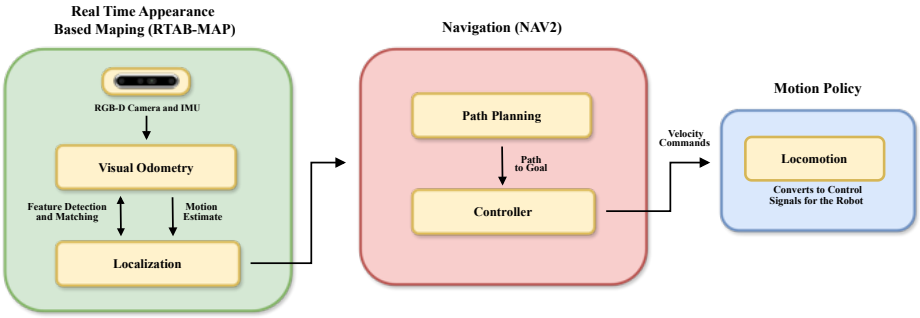


Fig. 5: Overview of the autonomous navigation stack.

program (project ref: https://doi.org/10.54499/2024.14126.UTA), titled FOMO-HODOR: FOundation MOdels for HumanOid DOmestic RObots[3].

The project focuses on three main research tasks:

- improving the use of Large Language Models (LLMs) in robot task planning by injecting common sense of LLMs into task planners.
- addressing the engineering challenges of integrating LLMs with a robot's AI architecture to enable effective human-robot interaction, with RoboCup@Home as a primary testing ground;
- enabling whole-body manipulation for humanoid robots.

Within RoboCup@Home, the General-Purpose Service Robot (GPSR) task requires robots to understand spoken commands, plan a sequence of action and execute the assigned task. Both UT Austin and ISR-Lisboa teams bring extensive experience with RoboCup@Home and GPSR, and are jointly developing a framework that combines prior work in robot architectures, knowledge representation, task planning and human-robot communication. This framework is being designed for the humanoid Booster T1 platform, equipped with grippers, and includes modules for locomotion, navigation, and vision-based grasping.

The team plans to demonstrate the outcomes of this collaboration at RoboCup@Home 2026, showcasing progress in humanoid robot autonomy within a domestic environment populated by humans.

References

1. Yuqian Jiang, Nick Walker, Minkyu Kim, Nicolas Brissonneau, Daniel S Brown, Justin W Hart, Scott Niekum, Luis Sentis, and Peter Stone. Laair: A layered architecture for autonomous interactive robots. In *Proceedings of the AAAI Fall Symposium on Reasoning and Learning in Real-World Systems for Long-Term Autonomy (LTA)*, October 2018.

[3] Project's web page: `https://irsgroup.isr.tecnico.ulisboa.pt/fomo-hodor/`

2. Yuqian Jiang, Nick Walker, Justin Hart, and Peter Stone. Open-world reasoning for service robots. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS 2019)*, July 2019.
3. Yoonwoo Kim, Raghav Arora, Roberto Martín-Martín, Peter Stone, Ben Abbatematteo, and Yoonchang Sung. Large-language-model-guided state estimation for partially observable task and motion planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2026. accepted.
4. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
5. Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. `https://github.com/snakers4/silero-vad`, 2024.
6. Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+p: Empowering large language models with optimal planning proficiency, 2023.
7. Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.
8. Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.
9. Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics (T-RO)*, 2023.
10. Adithyavairavan Murali, Balakumar Sundaralingam, Yu-Wei Chao, Jun Yamada, Wentao Yuan, Mark Carlson, Fabio Ramos, Stan Birchfield, Dieter Fox, and Clemens Eppner. Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training. *arXiv preprint arXiv:2507.13097*, 2025.
11. Balakumar Sundaralingam, Siva Kumar Sastry Hari, Adam Fishman, Caelan Garrett, Karl Van Wyk, Valts Blukis, Alexander Millane, Helen Oleynikova, Ankur Handa, Fabio Ramos, Nathan Ratliff, and Dieter Fox. Curobo: Parallelized collision-free robot motion generation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8112–8119, 2023.
12. Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. *arXiv preprint arXiv:2505.22642*, 2025.
13. Mathieu Labbe and Francois Michaud. Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29(3):734–745, 2013.
14. Mathieu Labbé and François Michaud. Long-term online multi-session graph-based splam with memory management. *Autonomous Robots*, 42(6):1133–1150, 2018.
15. Steve Macenski, Francisco Martín, Ruffin White, and Jonatan Ginés Clavero. The marathon 2: A navigation system. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2718–2725. IEEE, 2020.

Booster Software and External Devices [OPL]

We use a T1 humanoid from *Booster Robotics*. No modifications have been applied to the base hardware, but hands are replaced with EG2-4C2 electric gripper from InspireRobots for better compatibility with the current software stack.

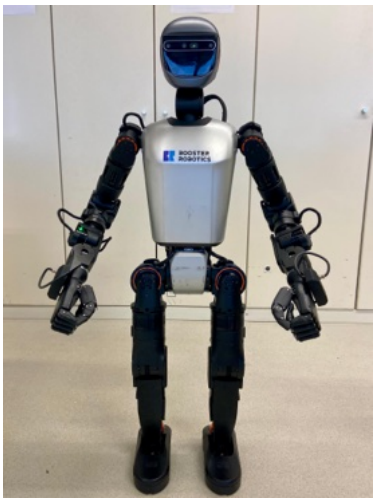


Fig. 6: Booster T1

Robot's Software Description

We are using the following 3rd party software:

- Object recognition: Grounded-SAM2
- People and activity recognition: Grounded-SAM2
- Grasp Sampler: GraspGen
- Knowledge Base: PostgreSQL
- Localization: RTAB-MAP
- Navigation: Nav2
- Speech Recognition: Whisper, Vosk
- Planning and reasoning: CoCo-TAMP
- State Machine: Yasmin (ROS2)

External Devices

No external device planned on being used.

Cloud Services

We are using the following cloud services:

- Speech recognition: Google Cloud Speech API
- Large language model: GPT-5