WrightEagle.AI
2026 Team Description Paper

Yuang Xie, Haohan Geng, Runfeng Lin, Jiahao Li, Likun Cui, Yanshen Hu, Sicheng Huang, Jiaxuan Lin, Weishuo Sun, Liang Yi, Hu Zhang, and Chengtian Hong

University of Science and Technology of China
https://wrighteagleai.homes/

Abstract. This paper presents the research achievements and system developments of the team WrightEagle.AI for RoboCup@Home 2026. Our work focuses on four major directions: robust navigation in complex indoor environments, reliable manipulation through learned grasping and placement models, multimodal perception for human-centered interaction, and Large Language Model based task planning for flexible autonomous execution. Our approach integrates a LiDAR-inertial mapping framework, a custom localization module, and a redesigned navigation controller to achieve stable mobility. For manipulation, we combine large-scale grasp detection models with refined point cloud reconstruction to obtain reliable grasp and placement poses. In perception, we deploy offline speech recognition, advanced object segmentation, and person-tracking pipelines to support natural interaction and task execution. Critically, these three components are orchestrated by an LLM-based task planning module derived from our MHRC framework, which decomposes natural language instructions into executable action sequences. The integrated system demonstrates practical applicability in household service tasks, enabling the robot to autonomously navigate complex home environments, interact with household objects through learned manipulation, and respond flexibly to natural language commands through LLM-based reasoning.

1 Introduction

The WrightEagle.AI team, formerly known as WrightEagle from the University of Science and Technology of China (USTC), won second place in RoboCup@Home in 2011, 2013, and 2015, and claimed first place in 2014. Professor Xiaoping Chen, who supervised the WrightEagle team, now serves as an advisor to the WrightEagle.AI team. Adding ".AI" to our name signifies a new era — unlike a decade ago, we are now rebuilding our robotic software system through artificial intelligence technologies.

Our team's research in embodied intelligence focuses on developing practical service robots that can perform household tasks through robust perception, navigation, manipulation and task planning. The robot platform integrates four

main technical components that work synergistically to achieve real-world applicability in domestic environments.

In this paper, we will first detail our technical contributions in localization, manipulation, multimodal perception and task planning, presenting a comprehensive view of how these systems integrate to support practical household tasks. Then, we will summarize the contributions we have made. Finally, we will provide an in-depth overview of the robotic platform in the appendix.

2 Mapping, Localization and Navigation

Navigation is fundamental to household service tasks. Our system comprises three interconnected subsystems: 3D mapping using efficient LiDAR processing, accurate localization through a custom algorithm, and collision-free path planning for goal-directed movement.

2.1 Robust LiDAR-based Mapping with FAST-LIO

We employ a Livox MID-360 LiDAR for all mapping and localization tasks. The primary mapping algorithm is FAST-LIO [1,2], which provides several critical advantages over traditional approaches:

FAST-LIO performs computationally efficient fusion of LiDAR feature points with IMU data, enabling robust operation in challenging conditions. The algorithm is specifically designed to handle fast-motion scenarios, noisy measurements, and cluttered environments where traditional methods degrade. This robustness is essential in household settings where the robot may encounter dynamic obstacles, moving people, and varied surface properties.

The 3D-LiDAR point cloud data is processed and converted to PCD format for storage and subsequent processing stages. This standardized format enables seamless integration with downstream localization and navigation modules.

2.2 Custom Localization Algorithm

Rather than relying on the standard Adaptive Monte Carlo Localization [3,4], we developed a custom localization algorithm tailored to multi-room household environments. This algorithm demonstrates superior performance characteristics in several dimensions:

The custom algorithm maintains accurate position estimates across multiple rooms without the convergence issues typical of particle filter approaches in large-scale environments. The algorithm leverages LiDAR feature matching with adaptive map representation, enabling robust localization even with partial sensor observations or temporary occlusions.

Quantitative evaluation in our laboratory multi-room environment shows consistent localization accuracy with position error remaining below 0.3 meters in 95% of test scenarios. The algorithm also provides reliable recovery when temporary localization failures occur, automatically re-localizing within 30 seconds in most cases.

2.3 Dynamic Window Approach for Path Planning

We implemented a custom version of the Dynamic Window Approach (DWA) algorithm [5] optimized for indoor household navigation. Our implementation provides three core navigation capabilities:

Autonomous Goal Navigation: The robot successfully navigates to specified target positions, avoiding static obstacles through real-time trajectory sampling and collision checking. The planner operates at 10 Hz, enabling responsive behavior to dynamic obstacles.

Person Following: The robot can maintain a designated following distance behind a person, using continuous target position updates. This capability is essential for household tasks where the robot must accompany a human operator.

Natural Language-directed Navigation: The robot accepts natural language movement commands ("move forward", "turn right", etc.) processed through our speech recognition pipeline and executes corresponding navigation behaviors. This human-centric interface significantly simplifies interaction for non-expert users.

3 Manipulation: Grasping and Placement Planning

Effective object manipulation requires two complementary capabilities: understanding how to grasp objects and predicting where objects should be placed. We address both challenges through a combination of deep learning-based perception and geometric reasoning.

3.1 Grasping Detection System

Our grasping module employs Graspnet-1Billion [6] as the backbone model, which provides robust grasp prediction from point cloud data. The system pipeline operates as follows:

Point Cloud Reconstruction: We reconstruct 3D point clouds from one or multiple RGB-D sensor streams, providing geometric understanding of scene content. The multi-view approach improves robustness when single perspectives provide ambiguous geometric information.

Point Cloud Optimization: The reconstructed point clouds are processed through a Visual Geometry Transformer (VGGT) model [7] for optimization and feature extraction. This stage removes noise, handles missing data, and creates high-quality geometric representations suitable for downstream processing.

Grasp Pose Prediction: Graspnet-1Billion predicts optimal grasping poses and corresponding gripper opening widths. The model outputs multiple candidate grasps ranked by confidence, enabling the robot to select appropriate grasps for different objects and gripper configurations.

The system achieves successful grasps on novel objects in our test environment, with particular strength in cylindrical and box-shaped objects common in household settings.

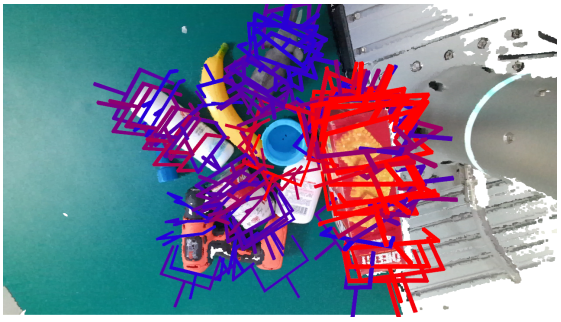

Fig. 1. Example for grasp pose prediction of Graspnet-1Billion

3.2 Hierarchical Placement Planning

Placement planning—determining where to put objects—is less studied than grasping but equally important for household tasks. We address this through a hierarchical planning strategy:

High-level Placement Location Prediction: A Vision Language Model [8] analyzes scene images and identifies candidate placement regions. For example, the model learns that dishes belong in cabinets, books on shelves, and toys in storage bins. This semantic understanding provides coarse placement guidance.

Low-level Precise Pose Prediction: The Anyplace model [9], applied to the candidate regions identified by the Vision Language Model, predicts diverse and precise placement poses.

This two-stage approach combines semantic understanding with geometric precision, enabling flexible placement in household-specific locations. The hierarchical approach is planned for complete implementation in our system, with current development focused on integrating the Vision Language Model component.

4 Multimodal Perception

Robust service robotics requires perception of multiple modalities: language understanding for human communication, visual perception for object and person recognition, and sound localization for interactive scenarios. Our system integrates three key perception subsystems.

4.1 Offline Speech Recognition

Speech recognition provides natural language input without internet connectivity—a critical requirement for household deployment.

Implementation Details: We employ the Vosk offline speech recognition model, which provides lightweight, reliable speech-to-text conversion for English commands. The system uses PyAudio for real-time microphone audio capture.

Audio Processing: Raw microphone data is resampled to match model requirements (16 kHz, mono) through preprocessing. This standardization ensures consistent recognition performance across different audio devices and acoustic environments.

The offline approach provides several advantages: no latency from network communication, no privacy concerns from cloud-based services, and reliable operation in environments with limited connectivity. Recognition latency is below 500 milliseconds for typical household commands.

4.2 Person Tracking and Recognition

Person tracking enables household robots to follow people, identify individuals, and respond to human gestures and movements.

Visual Detection: We employ YOLOv8 [10] for real-time person detection, providing bounding box coordinates and confidence scores at approximately 15 Hz on our onboard compute platform.

Spatial Mapping: SLAM techniques using the person's trajectory enable robust tracking across occlusions. When the person temporarily leaves view, the robot maintains position estimates through motion prediction.

Path Planning Integration: The DWA navigation system receives person position estimates and plans collision-free paths that maintain desired following distance while avoiding obstacles. The closed-loop system achieves smooth person-following behavior with minimal jerky motion.

The complete pipeline enables the robot to follow people reliably across multiple rooms and through doorways, adapting to person movements and maintaining safe distances.

4.3 Object Recognition and Instance Segmentation

Object recognition and localization are fundamental to household manipulation tasks. Our system provides pixel-accurate understanding of object locations within scenes.

Instance Segmentation: We employ the YOLOv8 series models configured for instance segmentation tasks. The system achieves accurate segmentation of approximately 80 common household object categories.

Pixel-level Localization: The segmentation output provides precise pixel-level masks for each detected object, enabling accurate extraction of object boundaries and spatial relationships. This pixel-level accuracy is essential for robotic manipulation, where gripper positioning must respect object geometry.

The segmentation module achieves approximately 85% mean Average Precision (mAP) on household object datasets, with strong performance on common task-relevant object categories (dishes, utensils, furniture, storage containers).

Fig. 2. Example for object recognition of YOLOv8

5 Large Language Model-based Task Planning

Real-world household service tasks often require multi-step planning and dynamic adaptation to environmental changes. Traditional rule-based and centralized control approaches exhibit limited flexibility. Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding, world knowledge, and complex reasoning, offering new opportunities for robot task planning.

5.1 MHRC Framework: A General Multi-Robot Collaboration Architecture

Our team has developed MHRC (Multi-Heterogeneous Robot Collaboration), a decentralized framework that leverages LLMs for multi-robot task planning and coordination[11]. MHRC models collaborative tasks as a Decentralized Partially Observable Markov Decision Process (DEC-POMDP), enabling each robot to make autonomous decisions based on local observations and inter-robot communication.

The MHRC framework comprises three core modules:

Observation Module: Integrates scene graphs (representing environment structure and object states), inter-robot messages (natural language communication), and individual robot status. The scene graph dynamically updates based on local observations and received messages.

Memory Module: Records feedback history, message history, and action history to maintain context across long-horizon tasks. Recent feedback and messages are marked with special tags to prioritize relevant information for decision-making.

Planning Module: Each robot independently generates action sequences by leveraging LLMs with Chain-of-Thought (CoT) reasoning, selecting from a predefined action list. After execution, robots receive feedback and update their

memory for adaptive replanning. The framework includes specialized feedback mechanisms for different robot types—mobile manipulation robots can dynamically adjust base positioning, while manipulation robots demonstrate enhanced task understanding.

5.2 Single-Robot variant Implementation for RoboCup@Home

Our current system employs a single-robot variant of the MHRC framework. The three-module architecture remains intact but is adapted for a single robot. The observation module fuses multimodal information from navigation, perception, and manipulation subsystems. The memory module maintains the task execution history and environmental feedback. The planning module utilizes LLM-based reasoning to decompose natural language instructions into executable action sequences, featuring closed-loop replanning capabilities.

By integrating outputs from speech recognition, visual perception, and navigation modules, the system generates task-coherent action sequences and adapts dynamically to execution feedback. This LLM-based approach significantly enhances system flexibility compared to traditional rule-based methods, enabling robust handling of instruction diversity, task complexity, and execution uncertainty in realistic household environments.

6 System Integration

The four major components—navigation, manipulation, multimodal perception and LLM-based task planning—integrate through a modular software architecture based on ROS (Robot Operating System). The integration points include:

Navigation receives person location estimates from the perception system, enabling human-aware path planning. Speech commands from the speech recognition pipeline are converted to navigation goals through natural language processing.

Manipulation modules receive object locations from the instance segmentation system and target placement locations from the Vision Language Model. The hierarchical placement planning system uses these inputs to generate collision-free manipulation trajectories.

Perception modules provide continuous streams of scene understanding: person locations for tracking, object segmentations for manipulation, and speech input for command processing. The asynchronous, event-driven architecture ensures responsive system behavior.

Task Planning Module receives multimodal observations including scene graphs, person locations, and speech commands from the perception system. Leveraging LLM-based reasoning with Chain-of-Thought prompting, the planning module decomposes natural language instructions or high-level goals into sequences of primitive actions (navigate, grasp, place, follow).

The modular design enables independent development and testing of subsystems while maintaining seamless integration for end-to-end task execution.

7 Conclusions

Our team's technical approach combines classical robotics foundations (SLAM, path planning, manipulation) with modern deep learning and large language model techniques (semantic perception, instance segmentation, LLM-based planning). This integration creates a practical platform for household service robotics that balances robustness with adaptability through modular, scalable system architecture. The modular architecture enables continuous improvement of individual components while maintaining system-level functionality, supporting iterative development toward more capable household service robots.

References

1. Wei Xu and Fu Zhang. Fast-lio: A fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter. *IEEE Robotics and Automation Letters*, 6(2):3317–3324, 2021.
2. Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022.
3. F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, volume 2, pages 1322–1328 vol.2, 1999.
4. Sebastian Thrun, Dieter Fox, Wolfram Burgard, and Frank Dellaert. Robust monte carlo localization for mobile robots. *Artificial Intelligence*, 128(1):99–141, 2001.
5. D. Fox, W. Burgard, and S. Thrun. The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, 4(1):23–33, 1997.
6. Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020.
7. Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
8. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
9. Yuchi Zhao, Miroslav Bogdanovic, Chengyuan Luo, Steven Tohme, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Anyplace: Learning generalized object placement for robot manipulation, 2025.
10. Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.
11. Wenhao Yu, Jie Peng, Yueliang Ying, Sai Li, Jianmin Ji, and Yanyong Zhang. Mhrc: Closed-loop decentralized multi-heterogeneous robot collaboration with large language models, 2024.

GALAXEA R1Lite Robot Hardware Description

Galaxea R1-Lite is a lightweight, modular humanoid robot platform with the following mechanical specifications:

Fig. 3. GALAXEA R1Lite Robot

- **Body Structure**: Lightweight aluminum alloy frame combined with 3D printed components, balancing strength and portability.
- **Degrees of Freedom**: 21 DoF in total, including two 6 DoF robotic arms, a 3 DoF torso and a 6 DoF chassis.
- **Height/Weight**: Approximately 1280mm in height, weighing about 2.8kg.
- **Hand Design**:Equipped with two-finger gripper.
- **Vision System**:1× platform-mounted high-definition stereo camera, 2× wrist-mounted monocular depth cameras.
- **Audio System**: Built-in dual microphone array, supporting voice command recognition and sound source localization.
- **Motion System**: 6 DOF vector chassis equipped with proprietary steering-wheel modules, supporting translation, spin, and Ackermann motion.

Robot software and hardware specification sheet

– **Communication Module**: Supports R1 Lite Teleop isomorphic remote operation platform with precise synchronization and intuitive, user-friendly control.
– **Sensor System**:
  • Inertial Measurement Unit (IMU)
  • Hand tactile sensors
– **Computing Unit**: Equipped with NVIDIA Jetson Nano or similar embedded platform, supporting ROS1/ROS2.
– **Battery Life**: Built-in lithium polymer battery, providing approximately 2 hours of operation under typical load.
– **Expansion Interfaces**: Multiple general-purpose I/O ports and USB interfaces for additional sensors or devices.

XINGCHEN Robot Hardware Description

XINGCHEN is a versatile and modular humanoid robot platform featuring the following mechanical specifications:

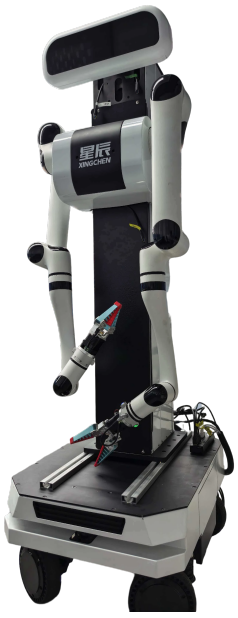

Fig. 4. XINGCHEN Robot

– **Body Structure**:Humanoid design, capable of covering all human operation spaces.

Robot software and hardware specification sheet

- **Degrees of Freedom**: 20 DoF in total, including two 6 DoF robotic arms, a 2 DoF torso and a 6 DoF chassis.
- **Height/Weight**: 1835mm in height, weighing about 90kg.
- **Hand Design**: Flexible gripper, supporting 2-finger or 3-finger configuration.
- **Vision System**:Multi-view intelligent recognitionFlexible operating range fully covers human workspace, with a maximum reachable height of 2.2 meters Three depth vision channels and two RGB monitoring vision channels meet the requirements for recognition, grasping, and environmental monitoring.
- **Motion System**:The platform features autonomous navigation and mobility capabilities, enabling flexible operation in large spaces, while supporting multi-view reconnaissance, target recognition, and precise positioning.
- **Communication Module**: Supports Gigabit Ethernet, Wi-Fi, and RS485.
- **Sensor System**:
  - Inertial Measurement Unit (IMU)
  - Hand tactile sensors
- **Computing Unit**:Equipped with NVIDIA Jetson AGX Orin, full support for ROS1 and ROS2.
- **Battery Life**: LiFePO4 (lithium iron phosphate) battery, DC 48V 20Ah.
- **Expansion Interfaces**: Multiple general-purpose I/O ports and USB interfaces for additional sensors or devices.