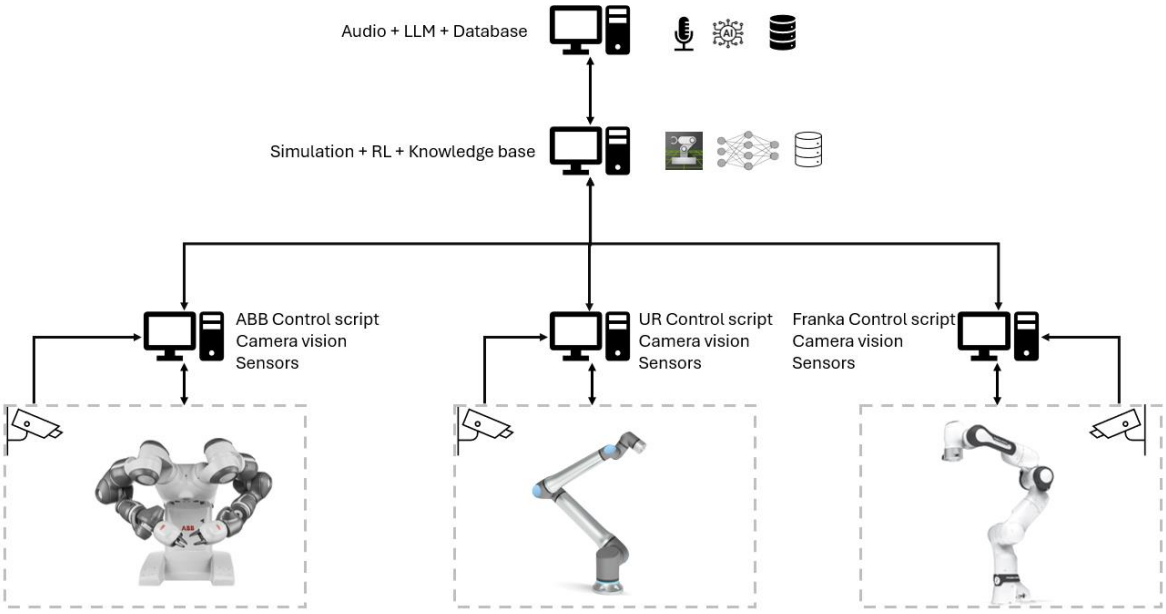
Integrating Vision, Language Models, and Robotic Control for Personalized Task Execution in Virtual Environments

Description:
This thesis proposes a unified framework that fuses state-of-the-art computer vision, large language models (LLMs), and advanced robotic control techniques to enable a robotic system that can naturally interact with users, understand their requests, and simulate task execution in a virtual environment. The ultimate goal is to establish an intelligent pipeline where a human can approach a collaborative robot - such as ABB's YuMi or a Universal Robots arm - issue a complex, high-level command in everyday language, and have the system seamlessly interpret, plan, and then virtually demonstrate the requested action.



To accomplish this, the research integrates four key components:

1. **Vision-Based Perception and User Recognition:**
   A robust computer vision pipeline will identify and recognize objects in the environment, detecting their spatial arrangements and affordances. In addition, face recognition and user identification methods will be employed to distinguish among multiple users. By linking recognized users to previously stored profiles in a secure database, the system can customize interactions, recall user preferences, and reference historical queries to provide a more natural and personalized experience.
2. **Language Understanding via LLMs:**
   The user's spoken or written instruction will be parsed by a large language model, which will translate the natural language request into a structured representation of tasks and subtasks. This component leverages state-of-the-art LLMs to resolve ambiguities, handle follow-up questions, and infer implicit requirements from context. Over time, the LLM will build conversational memory for each recognized user, enabling more natural and context-aware dialogues that adapt to the user's unique vocabulary and past interactions.

3. **Robotic Decision-Making and Control Schemes:**
   A decision-making engine, guided by the interpreted instructions from the LLM and informed by the visual scene understanding, will dynamically select appropriate task modules from a pre-defined skill library. These modules may range from simple pick-and-place maneuvers to more complex manipulations. The selected actions are then translated into control commands for the robot, ensuring they are feasible and safe. For the purpose of testing and validation, the initial focus will be on simulating these tasks with industrial robots like YuMi or a Universal Robots arm before deploying any learned behaviors in physical environments.
4. **Simulation in Omniverse for Validation and Refinement:**
   Before any command is physically executed, the entire scenario - environment layout, objects of interest, and the chosen robot decision tree - will be assembled in NVIDIA Omniverse. This virtual environment will serve as a testbed to refine the plan, adjust parameters, and validate the safety and efficiency of the proposed solution. By observing how the robot behaves in simulation, the framework can iteratively improve its understanding of the task and reduce errors prior to real-world deployment.

Intended Contributions:
This thesis aims to advance the state of human-robot interaction by providing a system that combines visual perception, language comprehension, and robot intelligence into a single, cohesive pipeline. By incorporating face recognition and historical user interaction data, it elevates the user experience, making dialogue more personal and efficient. The user can rely on a robot partner that remembers their name, understands their preferences, and quickly translates their natural language requests into tangible action plans.
The expected outcomes include:

- An end-to-end system demonstrated on a YuMi or Universal Robots platform, capable of taking user instructions and executing the corresponding tasks in simulation.
- A scalable approach to integrating vision and LLM-based language understanding for robotic command generation.
- A database-driven personalization layer, enabling human-robot dialogues to become more context-sensitive and user-centric over time.
- Empirical validation in Omniverse, ensuring that the system's simulated performance aligns with intended outcomes, improving reliability and reducing trial-and-error in physical testing.

By delivering a more intuitive, personalized, and verifiable approach to commanding robots, this research contributes to the growing field of human-robot collaboration and sets the stage for more complex, context-aware robotic systems in the future.

Contact:

Mehdi Tarkian; mehdi.tarkian@liu.se