Embodied AI

Learning to Grasp from Multiple Semantically Valid Options Using Reinforcement Learning and Language Guidance

Sergei Vedyshev[1], Eduardo Veas[2], Martina Mara[3], Belgin Mutlu[1]
Pro2Future GmbH[1], TUG-HCC[2], JKU-LIT Robopsychology Lab[3]
[1] Sandgasse 34, 8010 Graz, Austria
[2] Sandgasse 36, III, 8010 Graz, Austria
[3] Altenberger Straße 69, 4040 Linz, Austria

MOTIVATION & GOALS

Traditional industrial automation relies on rigid, pre-programmed systems that struggle with dynamic environments. Most current robotic grasping methods aim to find a single "best" grasp, failing if an object is blocked or slips. This project develops a highly adaptable **Embodied AI** framework for humanoid robots. By integrating **Large Language Models (LLMs)**, **Reinforcement Learning (RL)**, and **Causal Reasoning**, we aim to enable **Multi-Target Grasping** -- allowing robots to generate and choose from multiple viable grasp options based on semantic prompts. Furthermore, we focus on making these AI decisions **transparent and explainable**, fostering calibrated trust in **Human-Robot Collaboration (HRC)**.

Project FactBox

Project Name	Embodied AI
Project ID	StratP II A.S-1
Duration	12 Months

Area 3
Area Analytics

Project Lead
DI Dr. Belgin Mutlu

APPROACH

A closed-loop pipeline seamlessly bridges human demonstration and autonomous robot execution

- Capture complex human intent using **Motion Capture (CAMP)** and **VR Teleoperation**, collecting multi-modal data.
- Train a **Vision-Language-Action (VLA)** model using the collected data, beginning with offline learning and progressing to online **Reinforcement Learning (RL)** within high-fidelity physics simulation environments such as **NVIDIA Isaac Sim/Isaac Lab** to improve task performance and generalization.
- Deploy policies to the **G1 Robot** via **Sim-to-Real** continuously feeding real-world failures back into simulation to improve reasoning and grasping performance.

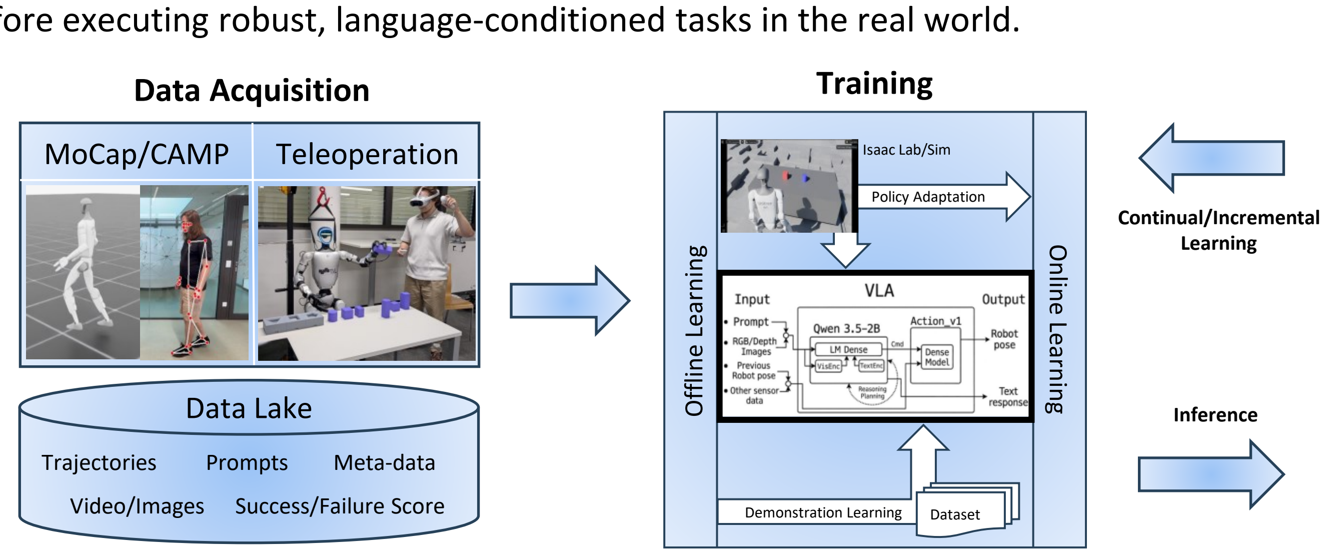
CONTRIBUTION

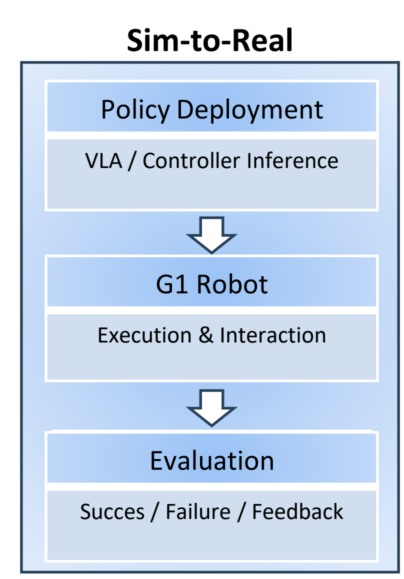
Scientific contribution
- Multi-Target Grasping & Adaptive RL: grasp selection via physical constraints and LLM prompts.
- Causal Reasoning: Structural models that predict grasp success and adapt to unexpected slips.
- Explainable AI (XAI): Real-time "think-aloud" verbalizations grounded in traceable data provenance.

Economic contribution
- Reduced Setup Costs: Replaces manual coding with intuitive human demonstrations and text commands.
- Increased Reliability: Adaptive grasping minimizes drop rates and errors in dynamic settings.
- Enhanced Worker Trust: Transparent, explainable actions accelerate safe adoption on the factory floor.
- Flexible Automation: Enables rapid retasking of robotic hardware for varying product lines.

SYSTEM ARCHITECTURE & VISUAL LAYOUT PROPOSAL

Multimodal human demonstrations are processed into a Data Lake, training a VLA model that practices in simulation before executing robust, language-conditioned tasks in the real world.





Contact: Sergei Vedyshev, Pro2Future GmbH, sergei.vedyshev@pro2future.at
Acknowledgement: This work was supported by Pro²Future II (FFG, 911655), TU Graz and JKU Linz.