Abstract

Artificial intelligence is now pervasive in our daily lives, from conversational agents and creative content generation to protein structure prediction and autonomous vehicles. As AI systems become increasingly integrated into the physical world, embodied AI, where intelligence is coupled with sensors and actuators to interact with the environment, has emerged as a key frontier, often regarded as the next industrial revolution with significant societal and economic impact. Despite impressive results, embodied agents are typically trained and evaluated under ideal benchmark settings, assuming perfect perception, clearly specified goals, or error-free, unambiguous language instruction. Yet, such assumptions rarely hold in real-world settings, where out-of-distribution conditions may occur and humans often provide underspecified or erroneous input. This thesis addresses these challenges by enhancing the robustness of embodied agents through active perception and human–robot interaction.

*To address **sensor unreliability**, we consider the problem of Active Visual Search (AVS), where the agent is tasked to find a specific object in an environment. In this setting, small objects, motion blur, low resolution, partial views, and heavy occlusions severely degrade both detection quality and planning accuracy. To this end, we present **POMP-BE-PD** [116], an unsupervised Monte-Carlo POMDP planner that explicitly models the uncertainty inherent in object detection to optimize the agent behavior. Specifically, the agent maintains a probability distribution on a 2D floor map, and incorporates the awareness that an object detector may fail into the aforementioned probability.*

*For the **instruction robustness**, we consider the fact that human instructions may be: (i) inaccurate or even prone to error, due to bad memory or cognitive/perceptual impairments; and (ii) ambiguous, as users may not be able or willing to supply all details of the task in advance. To this end, in [118] we first propose a novel task and benchmark of Detection and Localization of Instruction Errors, in which*

*different types of errors are injected into textual instructions. Moreover, we propose Instruction Error Detection & Localizer (**IEDL**) [118], a method to detect and localize errors within a sub-sentence distance on the original instruction. We extend this method in **I2EDL** [117], enabling agents to query users for clarification while minimizing their cognitive burden.*

*Finally, to handle **instruction ambiguity**, we introduce the Collaborative Instance object Navigation (**CoIN**) task [119], a new task setting where the agent actively resolves uncertainties about the target instance during navigation in natural, template-free, open-ended dialogues with humans. Our proposed agent, Agent-user Interaction with UncerTainty Awareness (**AIUTA**), leverages large Vision-and-Language models to reason about uncertainty and engage in efficient human-agent interaction.*