SECURE: SEMANTICS-AWARE EMBODIED CONVERSATION UNDER UNAWARENESS FOR LIFELONG ROBOT LEARNING

Rimvydas Rubavicius, Peter David Fagan, Alex Lascarides, Subramanian Ramamoorthy
School of Informatics
The University of Edinburgh
United Kingdom
{rimvydas.rubavicius,p.d.fagan,a.lascarides,s.ramamoorthy}@ed.ac.uk

ABSTRACT

This paper addresses a challenging interactive task learning scenario we call rearrangement under unawareness: an agent must manipulate a rigid-body environment without knowing a key concept necessary for solving the task and must learn about it during deployment. For example, the user may ask to "put the two granny smith apples inside the basket", but the agent cannot correctly identify which objects in the environment are "granny smith" as the agent has not been exposed to such a concept before. We introduce SECURE, an interactive task learning policy designed to tackle such scenarios. The unique feature of SECURE is its ability to enable agents to engage in semantic analysis when processing embodied conversations and making decisions. Through embodied conversation, a SECURE agent adjusts its deficient domain model by engaging in dialogue to identify and learn about previously unforeseen possibilities. The SECURE agent learns from the user's embodied corrective feedback when mistakes are made and strategically engages in dialogue to uncover useful information about novel concepts relevant to the task. These capabilities enable the SECURE agent to generalize to new tasks with the acquired knowledge. We demonstrate in the simulated Blocksworld and the real-world apple manipulation environments that the SECURE agent, which solves such rearrangements under unawareness, is more data-efficient than agents that do not engage in embodied conversation or semantic analysis.[1]

1 INTRODUCTION

A central theme in human-robot interaction (Bartneck et al., 2020) is to use embodied conversation (Cassell, 2001) to instruct and teach robots to perform a variety of tasks (Shridhar et al., 2021; Ahn et al., 2022; Brohan et al., 2023). In this pursuit, models for vision, language, and action have been developed to design generalist robotic agents (Shridhar et al., 2021; Ahn et al., 2022; Zitkovich et al., 2023; Octo Model Team et al., 2024). One challenge for these agents is the interactive task learning (ITL) scenario (Laird et al., 2017), in which the task instruction features a *neologism* (i.e., a newly coined expression) that denotes a concept the agent is unaware of but must understand to solve the task. For example, consider the scenario in figure 1. The user instructs the robot to "Put the two granny smith apples inside a basket". But before deployment, the robot was not trained on and cannot correctly ground the concept "granny smith", let alone distinguish objects that are "granny smith" from other objects, including other kinds of apple, such as "golden delicious". Achieving such a fine-grained categorisation (Wei et al., 2022) with limited prior experience might not be sufficiently captured in advance (Udandarao et al., 2024). For instance, Grounding DINO (Liu et al., 2023) can accurately detect a variety of objects in the environment, including an excellent ability to localize fruits. However, in our real-world experiments, grounding DINO does not work out-of-the-box when asked to localize objects denoted by the phrase "granny smith" or "granny smith apple", presumably because it completely lacked any examples labelled "granny smith" during training. To cope with this scenario, the agent must update their domain conceptualisation with this unforeseen concept, thereby expanding the set of possible states, which in turn demands revisions to planning about subsequent interactions with the user and the environment.

More generally, the task in figure 1 uses sensory observations to effect a rearrangement of rigid bodies into a state that satisfies the specified goal (Batra et al., 2020). In this work, we consider a special case of rearrangement, in which the goal is given in a natural language instruction and the agent is not aware of at least one concept required to understand it. We refer to such tasks as *rearrangement under unawareness*. To solve them, the agent can engage in an *embodied conversation* with the user to learn *online* how to recognise unforeseen concepts by interactively grounding them. This

[1]Project website: https://assistive-autonomy.github.io/secure

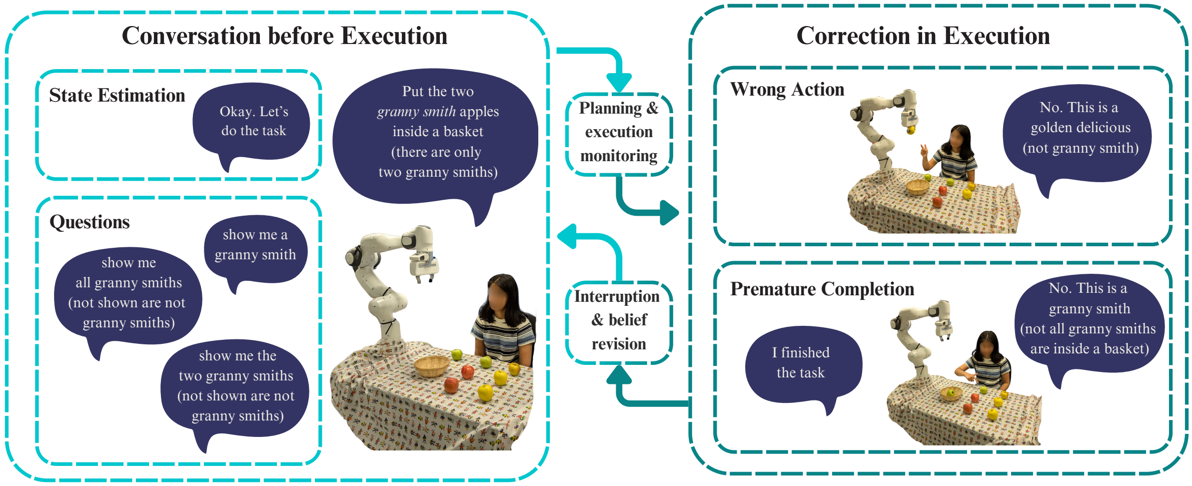


Figure 1: Framework overview. The agent interprets embodied conversation with the user in order to update its beliefs and solve tasks under unawareness: that is, an expression like "granny smith", which is a part of the user's instruction, is not in the agent's vocabulary, and the concept "granny smith" denotes is not a part of the agent's hypothesis space of possible domain models. Our framework enables the agent to exchange embodied messages with the user before attempting to solve the task; the expert can also provide feedback on the agent's actions in the environment during task execution. For conversation before execution, the agent issues questions (e.g., "show me a granny smith") to reduce its uncertainty about the domain. When executing the task, the user responds to a suboptimal action with embodied corrective feedback, e.g., "No. This is a golden delicious". Such corrections can occur in case of a wrong action (e.g., picking "golden delicious" instead of "granny smith") or premature completion (declaring that the goal state is reached when it is not.). Such feedback exposes the agent's false beliefs (being confident but wrong about the state). It triggers execution interruption and belief revision. In both cases, our framework processes embodied conversation in a semantics-aware manner: the user's messages are augmented with their logical consequences (shown in brackets in this figure). They are used to update the agent's beliefs, which in turn affects their decision-making.

requires incremental grounding (belief updates should occur with every user's message) and coping with an expanding hypothesis space (the dialogue may reveal unforeseen concepts). Consequently, the agent: (a) cannot assume mutual exclusivity between concepts; (b) must detect deficient beliefs about the hypothesis space; and (c) on discovering such deficiency, must adapt its prior beliefs to the newly expanded hypothesis space.

A principal way to process embodied conversation is to parse its utterances into a logical form in a symbolic logic that supports *automated inference* (Blackburn & Bos, 2005). This work studies how logical reasoning can enhance solving rearrangements under unawareness through interaction. We aim to make the agent *semantics-aware*: that is, besides the evidence acquired from directly processing the user's utterances, the agent also utilizes the *logical consequences* of the parsed output to aid learning and inference. Previous work showed the value of logical consequences in ITL stemming from the semantic analysis of individual messages (Rubavicius & Lascarides, 2022; Park et al., 2023) (e.g., "put the two granny smiths…" entails that there are two and only two objects that are "granny smith") as well as the semantic analysis of the overall discourse—in particular corrective feedback (Appelgren & Lascarides, 2020) (e.g., in figure 1, the user highlights a mistaken pick, which implies that the picked object is not "granny smith"). However, they don't *combine* sentence and discourse semantics and lack state-of-the-art perception and testing in the real world.

To rectify this gap in the existing work, we develop the SECURE ITL framework—**S**emantics-aware **E**mbodied **C**onversation under **U**nawareness for Lifelong **R**obot L**e**arning. SECURE combines symbolic reasoning and neural grounding to process the embodied conversation. The agent continuously updates its beliefs, and hence its plans, as and when the user provides information through their dialogue. It can discover unforeseen possibilities through the information exchange in the conversation, at which point SECURE revises its domain conceptualisation, triggering further belief updates and revised plans. SECURE also learns *how* to engage in conversation: that is, it learns how to make decisions that resolve the dilemma between asking questions (which comes at a cost) vs. risking solving the task using current beliefs, for which failure is very costly. Furthermore, our agent can process the user's embodied corrective feedback, in response to the mistakes it makes when attempting to solve the task: its interpretation of correction serves to revise (false) beliefs, and furthermore may compel it to engage in further conversation before taking further actions in the environment. To quantify the value of semantic awareness, we evaluate SECURE in a simulated blocksworld and in a real-world apple manipulation on several instances of rearrangement under unawareness. The results show that augmenting the users' messages with their logical consequences aids ITL.

2 FRAMEWORK

This section describes our ITL framework, including its components for knowledge representation and reasoning (section 2.1), interactive symbol grounding (section 2.2), and processing of embodied conversation (section 2.3).

2.1 KNOWLEDGE REPRESENTATION AND REASONING

The agent and the user exchange information using embodied utterances. These utterances consist of natural language expressions $\mathbb{N}$ and optional pointing gestures to objects $\mathbb{O}$ in the environment. Each natural language expression is mapped to its logical form $\mathbb{L}$ (see section 2.3), which supports automated model checking and inference.

SECURE uses a finite first-order domain model $\mathcal{M}$ (Gamut, 1991) to reason about the environment. $\mathcal{M}$ is a triple: a set of objects $\mathbb{O}$; a set of predicates, a.k.a. the vocabulary $\mathbb{V}$; and an interpretation function $I\colon \mathbb{V} \mapsto \mathbb{O}^*$ that maps predicates to denotations (for 1-place predicates it's a set of objects, for 2-place predicates a set of pairs of objects, and in general for n-place predicates a set of n-tuples of objects). The agent can evaluate statements in a natural language like "o_1 is a granny smith"$\in \mathbb{N}_{\text{snt}}$, which has the logical form $\texttt{grannysmith}(\texttt{o}_1) \in \mathbb{L}_{\text{snt}}$, to be true or false with respect to $\mathcal{M}$ using truth-conditional semantics. It can also identify objects denoted by various referential expressions $r \in \mathbb{N}_{\text{ref}}$: e.g., "the two granny smiths", which has the logical form $\Phi(r) = \langle \texttt{\_the\_2\_q}\ \texttt{x.grannysmith(x)} \rangle \in \mathbb{L}_{\text{ref}}$, is evaluated against $\mathcal{M}$ to yield either the empty set (reference failure) or a set consisting of the (unique) set of two objects in $\mathcal{M}$ that are in the range of $I(\texttt{grannysmith})$. We detail this (logical) reference semantics in appendix A. The agent uses the identified objects to understand the task instruction $t \in \mathbb{N}_{\text{task}}$ and, in turn, solve the task (e.g., from figure 1, when the two objects are identified as "granny smith", they are picked and placed in the basket.) The set of atomic predications (atoms) constructed from predicates followed by the appropriate number of constants that denote objects in the domain $\mathbb{O}$ is a set of ground atoms also known as Herbrand base $\mathbb{H}$. The set of atoms that are entailed by $\mathcal{M}$ is denoted as $\mathbb{H}_{\mathcal{M}} = \{\texttt{a} \in \mathbb{H} \mid \mathcal{M} \models \texttt{a}\}$.

The agent estimates the domain model $\hat{\mathcal{M}}$ using its belief state $b \in \mathbb{B}$ (Bochman, 2007; Mykel J. Kochenderfer & Wray, 2022). This maintains knowledge that the agent has acquired throughout its lifecycle from various sources of evidence, including percepts, memory, and priors. Percepts are constructed from the observations of the current situation. They include a set of objects in the environment $\mathbb{O}$, with corresponding object-centric embeddings $\boldsymbol{x} \in \mathbb{R}^d$ that are extracted from sensory observations in the environment and a domain theory Δ that is constructed (and updated) by processing the embodied conversation. Memory is a dynamic record of the agent's experience in its lifecycle. It includes the vocabulary $\mathbb{V}$ and support $\mathbb{S}$, which is a set of pairs consisting of an object-centric embedding $\boldsymbol{x} \in \mathbb{R}^d$ and the accompanying labels $\boldsymbol{y} \in [0,1]^{|\mathbb{V}|}$ with $y^{(\texttt{p})}$, indicating the degree of belief that an object with the corresponding embedding is denoted by the symbol $\texttt{p} \in \mathbb{V}$. Priors that determine the initial beliefs of each atom are given by prior weights $w_p\colon \mathbb{H} \mapsto [0,1]$, parameterizing the Bernoulli random variable Bern for each atom. They can come from the initial estimation based on sensory observations (as in figure 1) or, in the case of pure unawareness, they are assigned 0.5 as the *default weight*, to capture complete ignorance about denotations of the symbol in question. Using the belief state, the agent can reason about the domain and make decisions. To do this, evidence is combined into *grounded weights* $w_g\colon \mathbb{H} \mapsto [0,1]$ that parameterise a probability distribution over domain models. In particular, using weights (both w_p and w_g), a probability of a well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$ can be computed using weighted model counting WMC (Chavira & Darwiche, 2008) that counts and weights events (instances of $\mathcal{M}$) that are valid under constraints ($\mathcal{M}$ entails ϕ):

$$\mathrm{WMC}_w(\phi) = \sum_{\mathcal{M}\colon\, \mathcal{M} \models \phi} \prod_{\texttt{a}\colon\, \texttt{a} \in \mathbb{H}} w(\texttt{a}) \cdot \mathbf{1}_{\texttt{a} \in \mathbb{H}_{\mathcal{M}}} + (1 - w(\texttt{a})) \cdot (1 - \mathbf{1}_{\texttt{a} \in \mathbb{H}_{\mathcal{M}}}) \tag{1}$$

WMC is used to compute various (probabilistic) queries given the current belief state, such as the conditional probability CON of a formula ϕ, and the most probable domain model (MAP estimate):

$$\mathrm{CON}_w(\phi \mid b) = \frac{\mathrm{WMC}_w(\phi \wedge \tilde{\Delta})}{\mathrm{WMC}_w(\tilde{\Delta})} \quad \hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \mathrm{CON}_w(\mathcal{M} \mid b) \tag{2}$$

where $\tilde{\Delta}$ is a conjunction of formulas in the domain theory Δ of the belief state b and $\mathrm{CON}_w(\mathcal{M} \mid b)$ is the probability of $\mathcal{M}$ given by the conditional probability of the conjunction of model atoms based on them being in $\mathbb{H}_{\mathcal{M}}$, or not.

2.2 INTERACTIVE SYMBOL GROUNDING

The grounding model is a function from sensory observations to atoms that hold in the domain.[2] It is operationalised as a probabilistic classifier $\omega_b \colon \mathbb{R}^d \mapsto [0,1]^{|\mathbb{V}|}$ that depends on the belief state $b \in \mathbb{B}$. Given an embedding $\boldsymbol{x} \in \mathbb{R}^d$, corresponding to an object $o \in \mathbb{O}$, which is extracted from the environment's sensory observations, ω_b predicts binary labels $\hat{\boldsymbol{y}} \in [0,1]^{|\mathbb{V}|}$ with each element $\hat{y}^{(\mathtt{p})}$ indicating a probability that an atom $\mathtt{p}(\mathtt{o})$ constructed from predicate $\mathtt{p}$ and the constant $\mathtt{o}$ that denotes the object o with an embedding $\boldsymbol{x}$ is a member of $\mathbb{H}_{\mathcal{M}}$ (equivalent to the probability that ground-truth domain model $\mathcal{M}$ entails $\mathtt{p}(\mathtt{o})$ based on the current belief state and object-centric embedding):

$$\hat{\boldsymbol{y}} = \omega_b(\boldsymbol{x}) \quad \hat{y}^{(\mathtt{p})} = P(\mathtt{p}(\mathtt{o}) \in \mathbb{H}_{\mathcal{M}} \mid b)$$

For interactive symbol grounding, SECURE uses a non-parametric grounding model of multilabel prototype networks (Yang et al., 2019; Cano Santín et al., 2020). This uses the support $\mathbb{S}$ of the belief state b to make predictions. The unique approach of SECURE lies in using the domain theory Δ, which is built by processing the latest user's (embodied) utterance to update $\mathbb{S}$ and, in turn, update the belief state. In particular, whenever the agent constructs a new well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$, the belief updates $\mathrm{Update} \colon \mathbb{B} \times \mathbb{L} \mapsto \mathbb{B}$ with new support constructed by updating the label values with the new conditional probabilities with respect to prior weights $\mathrm{CON}_{w_p}(\cdot \mid b)$ (see Appendix B for details).

Using support $\mathbb{S}$, the grounding model builds positive and negative support for each symbol $\mathtt{p} \in \mathbb{V}$, denoted as $\mathbb{S}^{(+\mathtt{p})}$ and $\mathbb{S}^{(-\mathtt{p})}$, respectively. Whether a pair goes into the support is decided by: the value of $y^{(\mathtt{p})}$, the entropy of $\mathrm{Bern}(\mathrm{y}^{(\mathtt{p})})$, and the threshold τ for the overall noise level in support building:

$$\mathbb{S}^{(+\mathtt{p})} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{S} \mid y^{(\mathtt{p})} > \frac{1}{2} \wedge H(\mathrm{Bern}(\mathrm{y}^{(\mathtt{p})})) \leq \tau\} \qquad \mathbb{S}^{(-\mathtt{p})} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{S} \mid y^{(\mathtt{p})} < \frac{1}{2} \wedge H(\mathrm{Bern}(\mathrm{y}^{(\mathtt{p})})) \leq \tau\}$$

Using $\mathbb{S}^{(+\mathtt{p})}$ and $\mathbb{S}^{(-\mathtt{p})}$, positive and negative prototypes, denoted as $\boldsymbol{z}^{(+\mathtt{p})}$ and $\boldsymbol{z}^{(-\mathtt{p})}$ respectively, are computed by taking the weighted average of embeddings of the corresponding supports with the weight being the likelihood that the relevant object is (or respectively is not) denoted by $\mathtt{p}$ (i.e., the values $y^{(\mathtt{p})}$ and $(1 - y^{(\mathtt{p})})$ respectively):

$$\boldsymbol{z}^{(+\mathtt{p})} = \frac{1}{|\mathbb{S}^{(+\mathtt{p})}|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{S}^{(+\mathtt{p})}} y^{(\mathtt{p})} \boldsymbol{x} \qquad \boldsymbol{z}^{(-\mathtt{p})} = \frac{1}{|\mathbb{S}^{(-\mathtt{p})}|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{S}^{(-\mathtt{p})}} (1 - y^{(\mathtt{p})}) \boldsymbol{x}$$

If, due to a lack of evidence (so far), the positive/negative support is empty—there is not enough evidence (yet) to deem any of the observed embeddings to be *good enough*—then positive/negative prototypes default to support elements with the largest/smallest entropy as the best guess of what exemplars are suitable for positive/negative support.

Using positive/negative prototypes, the probability that predicate $\mathtt{p}$ is true of an object with embedding $\boldsymbol{x}$ is computed using their cosine distance as an activation for a sigmoid: $\hat{y}^{(\mathtt{p})} = \sigma(\cos(\boldsymbol{z}^{(-\mathtt{p})} - \boldsymbol{z}^{(+\mathtt{p})}, \boldsymbol{x}))$. Grounding model predictions yield grounded weights $w_g(\mathtt{p}(\mathtt{o})) \leftarrow \omega_b(\boldsymbol{x})^{(\mathtt{p})}$ that can be used for reasoning (section 2.1).

The presented grounding model handles tasks under unawareness in the following way. Suppose that the agent observes a new symbol or neologism $\mathtt{p}^*$ from the latest user's message. The agent becomes aware of $\mathtt{p}^*$ by extending its vocabulary $\mathbb{V} \leftarrow \{\mathtt{p}^*\} \cup \mathbb{V}$ in the belief state and by updating the belief state—in particular, the dimensionality of $\boldsymbol{y}$ increases by populating it with prior weights w_p, which in case of pure unawareness is the default weight 0.5.

2.3 PROCESSING EMBODIED CONVERSATION

In processing embodied conversation, the agent performs semantic analysis (section 2.3.1), which contributes to decisions about action before attempting the task (section 2.3.2) and to understanding the user's corrections (section 2.3.3).

2.3.1 SEMANTIC ANALYSIS

When the user utters an embodied signal, its semantic analysis yields a well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$, which gets added to Δ and triggers a belief update. Semantic analyses include: sentence-level $\xi \colon \mathbb{N}_{\text{ref}} \times 2^{\mathbb{O}} \mapsto \mathbb{L}_{\text{snt}}$, stemming from generalised quantifier theory (**?**); and discourse-level $\zeta \colon \mathbb{N}_{\text{task}} \times \mathbb{N}_{\text{corr}} \times \mathbb{A} \mapsto \mathbb{L}_{\text{snt}}$, stemming from coherence-based theories of discourse, in particular coherent corrections (Lascarides & Asher, 2009).

[2] We only consider one-place predicate grounding here. In principle, the grounding model could be extended to predicates of arbitrary arity. For instance, for two-place predicates involving spatial relationships, we could consider embeddings constructed from the pair of embeddings of the objects participating in this relationship.

The sentence-level semantic analysis $\phi \in \mathbb{L}_{\text{snt}}$ is constructed from a referential expression $r \in \mathbb{N}_{\text{ref}}$ and objects in the environment denoting it known as the referent $\mathcal{R} \subseteq 2^{\mathbb{O}}$. This is a subset of $2^{\mathbb{O}}$, with each element being a set of objects denoted by the referential expression. E.g., when 3 objects are "granny smith" the referent of "two granny smiths" is a set of 3 elements, each containing 2 out of the 3 objects that are "granny smith". ϕ includes clauses deduced from the surface form of r as well as negated clauses from the logical consequence of using a quantifier (Beaver, 1997). For instance, if the user utters the referential expression $r =$"the one granny smith" and points to o, this leads to the formula $\xi(r, \mathcal{R}) = \texttt{grannysmith(o)} \wedge \bigwedge_{o' \in \mathbb{O}-o} \neg\texttt{grannysmith(o}')$ getting added to Δ. The deduced negated clauses follow from the uniqueness condition on the quantifier "the one" (Russell, 1917), which entails that no other object in the environment is denoted by the phrase "granny smith".

For discourse-level semantic analysis, following Lascarides & Asher (2009), the user's correction $c \in \mathbb{N}_{\text{corr}}$ in response to the agent's action $a \in \mathbb{A}$ has two effects on the agent's beliefs: 1) the content of $c \in \mathbb{N}_{\text{corr}}$ is true; and 2) $a \in \mathbb{A}$ is not part of an optimal plan, which leads to further deductions that are dependent on the task $t \in \mathbb{N}_{\text{task}}$ and what type of action a is. Section 2.3.3 looks in detail at which well-formed formulae are inferred and added to Δ.

2.3.2 Conversation before Execution

When solving a task under unawareness, the interaction begins with the user uttering a task instruction in natural language $t \in \mathbb{N}_{\text{task}}$. Before trying to solve the task using their (current) belief, this option being a sequence of a_{act} from the set of possible executable actions $\mathbb{A}_{\text{exec}}$ that change the environment, the agent can choose instead to ask a question $a_{\text{quest}} \in \mathbb{A}_{\text{quest}} \subset \mathbb{A}$, to reduce uncertainty about the domain model $\hat{\mathcal{M}}$. The questions are of the form "show me r" where $r \in \mathbb{N}_{\text{ref}}$ is a referential expression. The user responds to such questions by pointing to a referent or r, which, using sentence-level semantic analysis ξ, yields a well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$ that triggers a belief update. ϕ is not known in advance to the agent and can be conceptualized as a random variable parameterized by the agent's $\phi \sim \text{Result}(a)$. To maintain coherence, $\mathbb{A}_{\text{quest}}$ depends on the task instruction: specifically, r in the question can feature only non-logical words used in the instruction, and its quantifier can be replaced with an existential or universal. For example, in the case of $t =$"move every red cylinder to the left of the one cube", the agent can ask to show "every red cylinder", "the one cube", "a red cylinder", "a cube" and "all cubes", but not for instance "a red cube" as this referential expression may not have a referent in the environment.

To make strategic decisions about whether and how to engage in the embodied conversation, SECURE solves the following decision problem. Each agent's action ($\mathbb{A} = \mathbb{A}_{\text{quest}} \cup \mathbb{A}_{\text{exec}}$) results in a well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$ being observed, followed by the belief update $\text{Update}(b, \phi)$. To quantify the value of each action, we take an information-theoretic approach and measure the *expected information gain* $I \colon \mathbb{B} \times \mathbb{A} \mapsto \mathbb{R}^{+}$ (Lindley, 1956; Thrun et al., 2005) for actions:

$$I(b, a) = H(b) - \mathbb{E}_{\phi \sim \text{Result}(a)}[H(\text{Update}(b, \phi))] \tag{3}$$

The entropoy of belief states is defined using grounded weights w_g to parameterise the Bern random variable for each atom:

$$H(b) \coloneqq \sum_{\mathtt{a} \in \mathbb{H}} H(\mathrm{b}) \quad \mathrm{b} \sim \text{Bern}(w_g(\mathtt{a})) \tag{4}$$

The expectation in equation 3 is computed by marginalising over answers ϕ consistent with b, with the likelihood that ϕ is the actual answer computed using conditional probabilities with grounded weights (equation 2).

The value of information, of performing certain actions in the environment is in a trade-off with rewards of taking these actions. Questions have an inherent cost $C \colon \mathbb{A}_{\text{quest}} \mapsto \mathbb{R}^{+}$, which approximates the user's answering effort. It depends on two quantities: $\text{Obj}(r)$ that is the number of objects in the referent of r approximating the designation (pointing) effort, and $\text{Sym}(r)$ that is the number of predicates in $\Phi(r)$, approximating reference resolution effort. These quantities are weighted by unit pointing cost C_{point} and unit reference cost C_{ref}, resulting in the overall cost:

$$C(q) = C_{\text{point}}\text{Obj}(r) + C_{\text{ref}}\text{Sym}(r) \quad r \in a_{\text{quest}} \tag{5}$$

For example, question $q =$"show me the one red cube" cost is $C_{\text{point}} + 2C_{\text{ref}}$ as its referential expression ($r =$"the one red cube") features two predicates and answering involves pointing to one object.

For action $a_{\text{act}} \in \mathbb{A}$ and successful attempt to solve the task the reward of 1 is observed and -1 otherwise, resulting to the following reward function $R \colon \mathbb{A} \times \mathbb{B} \mapsto [-1, 1]$:

$$R(a) = \begin{cases} 1 & \text{if } a = a_{\text{act}} \text{ and task is solved correctly} \\ -C(a) & \text{if } a \in \mathbb{A}_{\text{quest}} \\ -1 & \text{otherwise} \end{cases} \tag{6}$$

SECURE compares expected information gain to the expected reward of the current belief state:

$$\mathbb{E}_b[R(a)] = \begin{cases} -\tilde{C}(a) & \text{if } a \in \mathbb{A}_{\text{quest}} \\ 2 \cdot \text{CON}_{w_g}(\mathcal{M} \mid b) - 1 & \text{otherwise} \end{cases} \tag{7}$$

where $\tilde{C}$ is just like C except different for the universal quantifier as the number of objects in a referent is not known in advance as it ranges from 1 to $|\mathbb{O}|$, thus is approximated by the average, i.e., $\frac{|\mathbb{O}|-1}{2}$.

This leads to the overall action-value function $Q\colon \mathbb{B} \times \mathbb{A} \mapsto \mathbb{R}$:

$$Q(b, a) = \theta_1 I(b, a) + \theta_2 \mathbb{E}_b[R(a)] = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \cdot \begin{bmatrix} I(b, a) \\ \mathbb{E}_b[R(a)] \end{bmatrix} = \theta^\top h(b, a) \tag{8}$$

where $\theta = [\theta_1, \theta_2]^\top$ are parameters signifying preferences in exploration (engaging in embodied conversation) and exploitation (solving task under unawareness) and $h(a, b) := [I(b, a), \mathbb{E}_b[R(a)]]^\top$ known as preference (Sutton & Barto, 1998). Action-value function Q parameters θ can be optimized using semi-gradient SARSA. See appendix B for details.

2.3.3 Correction in Execution

Upon choosing $a_{\text{act}} \in \mathbb{A}$, planning with $\hat{\mathcal{M}}$ is used to compute the sequence of execution actions. For rearrangement, three execution actions $\mathbb{A}_{\text{exec}} \subset \mathbb{A}$ are considered: a_{pick}, for picking an object in the environment, a_{place} for placing the picked object in the target location and a_{complete} for declaring that the agent has solved the task.

When executing these actions, the agent can make mistakes (suboptimal actions in the environment) due to an incorrectly estimated domain model $\hat{\mathcal{M}}$. The user reacts to a mistake by providing embodied corrective feedback. Each correction consists of the corrective cue "No.", followed by a referential expression and designation that highlights the source of error, e.g. "No. This is a cylinder. (points to $o_{\text{corr}} \in \mathbb{O}$)". As mentioned in section 2.3.1, semantic analysis of the correction depends on the task instruction $t \in \mathbb{N}_{\text{task}}$, the executed action $a \in \mathbb{A}_{\text{exec}}$, and the correction itself $c \in \mathbb{N}_{\text{corr}}$. To illustrate the different discourse-level semantic analyses, we consider the task instruction $t =$"move every cube in front of a cylinder".

When the action $a_{\text{pick}} \in \mathbb{A}_{\text{exec}}$ gets corrected, semantic analysis dictates that the picked object should *not* have been picked: specifically, the picked object is not denoted by the direct object referential expression r_{direct} in the task instruction $t \in \mathbb{N}_{\text{task}}$. For instance, when pick is corrected with $c =$"No. This is a cylinder. (points to $o_{\text{corr}} \in \mathbb{O}$)", semantic analysis dictates that $\texttt{cylinder}(\texttt{o}_{\text{corr}})$ and $\neg\texttt{cube}(\texttt{o}_{\text{corr}})$ are both true. In general, the semantic analysis for pick is as follows:

$$\zeta(a_{\text{pick}}, t, c) = \xi(r_{\text{corr}}, \{\{o_{\text{corr}}\}\}) \wedge \neg\xi(r'_{\text{direct}}, \{\{o_{\text{corr}}\}\})$$

where r'_{direct} is just like r_{direct} but with an existential quantifier.

When the action $a_{\text{place}} \in \mathbb{A}_{\text{exec}}$ gets corrected, semantic analysis is similar to the `pick` case, except this time the object designated in the correction indicates that it is not part of the referent of the indirect object referential expression r_{indirect} in t. In our illustrative example, if the correction $c =$"No. This is a sphere. (points to $o_{\text{corr}} \in \mathbb{O}$)" is produced in response to $a_{\text{place}} \in \mathbb{A}_{\text{exec}}$, then the agent would (again) be able to infer $\texttt{sphere}(\texttt{o}_{\text{corr}})$ directly from correction. But this time, the agent would also infer $\neg\texttt{cylinder}(\texttt{o}_{\text{corr}})$ because placing the object must result in the placed object being in the intended spatial relation to the indirect object. In general, the semantic analysis for place is as follows:

$$\zeta(a_{\text{place}}, t, c) = \xi(r_{\text{corr}}, \{\{o_{\text{corr}}\}\}) \wedge \neg\xi(r'_{\text{indirect}}, \{\{o_{\text{corr}}\}\})$$

where r'_{indirect} is a referential expression just like r_{indirect} but with a existential quantifier.

When the action $a_{\text{complete}} \in \mathbb{A}_{\text{exec}}$ gets corrected, the correction entails that the goal state is *not* reached. To illustrate, consider correction $c =$"No. This is a cube (points to $o_{\text{corr}} \in \mathbb{O}$)" utterend in response to complete. The formula $\texttt{cube}(\texttt{o}_{\text{corr}})$ is a part of the logical form of this message. So the agent must reject any domain model for which $\neg\texttt{cube}(\texttt{o}_{\text{corr}})$ is true, including its previous MAP estimate. Furthermore, the fact that this logical form is a correction entails that the current state does not satisfy the goal: in other words, there is a cube that is *not* in front of a cylinder. Treating the message "this is a cube (points to $o_{\text{corr}} \in \mathbb{O}$)" as an *explanation* of the corrective cue "No" entails that it is o_{corr} that is the renegade cube, not in front of a cylinder.[3] In general, the semantic analysis for complete is as follows:

$$\zeta(a_{\text{complete}}, t, c) = \xi(r_{\text{corr}}, \{\{o_{\text{corr}}\}\})$$

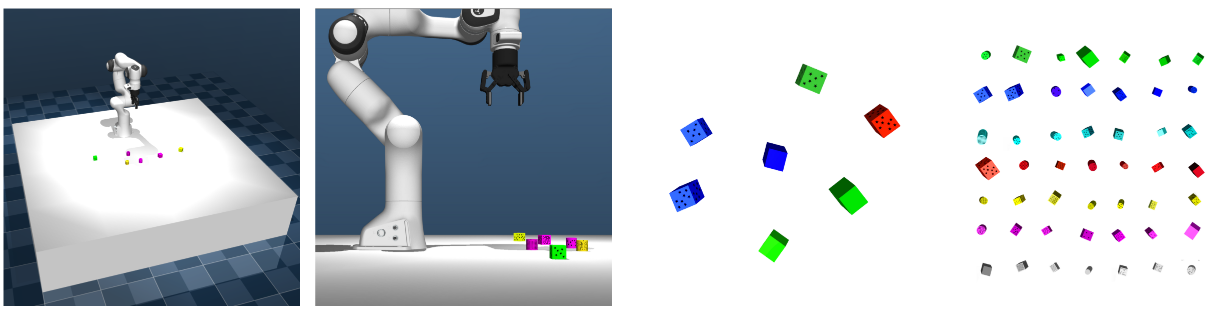
[3]With the ground truth to spatial relations, the agent could infer from this that every object in front of o_{corr} is not a cylinder. SECURE foregos exploiting this entailment, however, for the sake of computational efficiency.

3 EXPERIMENTS

The proposed ITL framework is evaluated in simulation and real-world experiments on a Franka Emika Panda robot. We compare SECURE's policy π_{secure} against two baselines π_{simple} and π_{correct}, which are as follows:

- π_{secure}: the agent can engage in embodied conversation using semantic analysis to reason about both the literal interpretation and logical consequences of different actions. It can also process correction in execution;
- π_{simple}: the agent can engage in conversation before execution but does not compute any logical consequences: questions "show me the one cube" and "show me a cube" have the same expected information gain, and the same belief state after the user's response. It can also process correction in execution;
- π_{correct}: the agent cannot engage in embodied conversation but processes correction in execution.

Each policy performance is measured using three performance metrics: cumulative reward cR and commutative cost cC as an extrinsic metric of task success as well ass mean micro F1 score $\mathrm{mF1}$ between ground truth and estimated referents at the beginning of each task for symbols in the task instruction, as an intrinsic metric for grounding. See appendix C for implementation details.



(a) Bird-eye view (b) Side view (c) Top-down view (d) Object variability

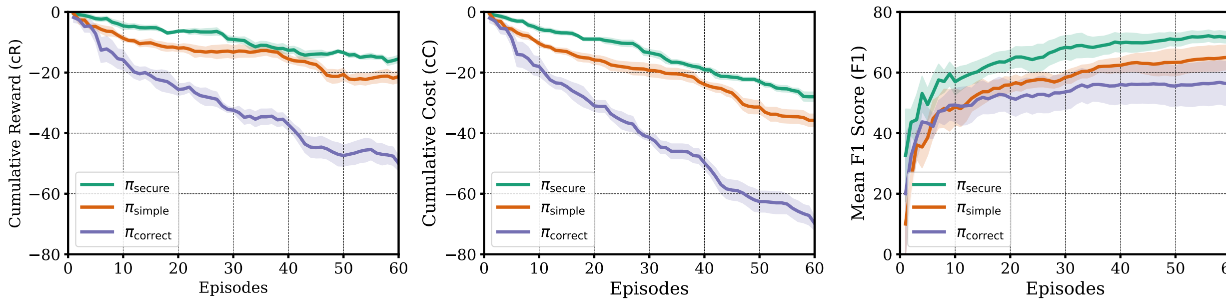
Figure 2: Overview of MuJoCo (Todorov et al., 2012) environment used for simulation experiments (section 3.1) including different views (see figures 2a, 2b, and 2c) and overview of object variability (see figure 2d). Each environment contains 6-7 rigid body objects. Each task instruction uses concepts of texture (plain, dotted, start), shape (cube, rectangle, cylinder), and colours (red, green, blue, cyan, grey, magenta, yellow) that agent is not aware of before the deployment as well predefined concepts of spatial relationships (in front of, to the right of, to the left of, behind) and quantifiers (a/an, every/all, and the n, where n is natural number).

3.1 SIMULATION EXPERIMENTS

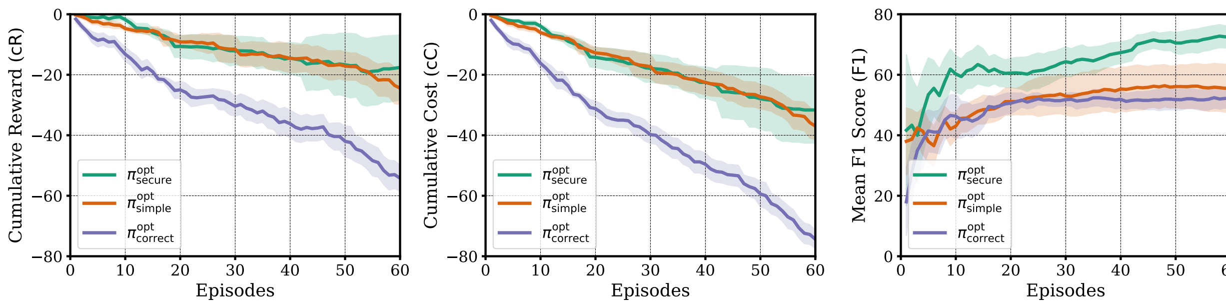
We conduct blocksworld manipulation experiments in simulation (see figure 2). Each task instruction has the form "move r_1 rel r_2" where rel is a spatial relationship and $r_{1,2}$ are referential expressions consisting of a quantifier followed by one to three symbols (e.g. "every red cube"). CodeLlama7B (Rozière et al., 2023) is used for semantic parsing to obtain logical forms of referential expressions and DINOv2 (Oquab et al., 2023) for feature extraction to obtain object-centric embeddings. The environment includes an oracle: it answers the agent's questions and corrects its suboptimal actions (see section 2.3.3). Each agent attempts to solve the task up to 5 times (one episode), with their policy evaluated over the same sequence of 60 episodes, averaged over 5 runs.

3.1.1 ITL EXPERIMENT

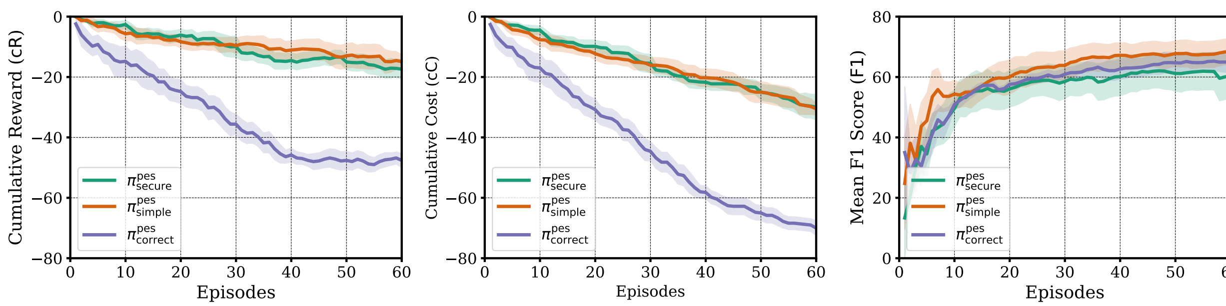
The results of the simulation experiment are given in figure 3a. The cumulative reward cR and commutative cost cC are taking negative values, indicating both the agent's questions and the user's corrections before succeeding (or not) in executing the task. Note that the cR curve is steeper than cC due to the positive rewards received from solving the task. Agents that engage in conversation before execution (i.e., π_{secure} and π_{simple}) receive a higher and statistically different cR and cC compared with π_{correct} (paired t-test p-values: π_{secure} v.s. π_{correct} : $4 \cdot 10^{-26}$ and π_{simple} v.s. π_{correct} : $4 \cdot 10^{-8}$). This signifies the value of the learner having some control over the information they receive from the conversation before attempting to solve the task — a form of active learning. Comparing cR for π_{secure} and π_{simple}, we observe that for the majority of the tasks π_{secure} performs better than π_{simple} except for a brief task period 33-37 in which there's no statistically significant difference, indicating that both policies have similar



(a) ITL experiment



(b) Optimistic belief experiment



(c) Pessimistic belief experiment

Figure 3: Simulation experiment results. Learning curves are given with 95% confidence interval.

embodied conversations and interactions in the environment. However, the $\mathrm{mF1}$ of π_{secure} performs on average 5% better than π_{simple} and 10% better than π_{correct}. This signifies that it is beneficial to use agents whose decision-making and processing utilise the logical consequences of the user's embodied conversation messages.

3.1.2 Optimistic and Pessimistic Belief Experiments

We experiment with SECURE's ability to cope with (initial) biased beliefs that might be false. To evaluate this, we consider a variation on the experiments from section 3.1.1, in which all textures (plain, dotted, starry) are assigned optimistic $w_p^{\text{opt}} = 0.7$ and pessimistic $w_p^{\text{pes}} = 0.3$ prior weights: these respectively impose an initial bias that all objects are (and respectively are not) plain, dotted and starry. Such bias encourages (respectively discourages) symbol prediction. The results are in figures 3b and 3c. When comparing cR and cC for agents with biased beliefs, π_{correct} performs worse than agents that engage in embodied conversation. For π_{simple} and π_{secure}, for the period of the first 20 tasks, π_{secure} performs better under biased beliefs while later the performance difference is not conclusive, fluctuating between π_{simple} or π_{secure}. The $\mathrm{mF1}$ scores show that w_p^{opt} hurts π_{simple}, making its grounding performance similar to π_{correct}, while π_{secure}'s grounding performance is similar to that in section 3.1.1. On the other hand, w_p^{pes} does not yield any statistically significant difference among the agents. This is not surprising: the value derived from the semantic analysis diminishes because it predicts that atoms are false, which aligns with the pessimistic bias. We observe that $\pi^{\text{opt}}_{\text{secure}}$ has high variance due to diverse dialogues, involving both attempts to solve the task without dialogue and questions that exploit semantic analysis, leading to more diverse belief states.

3.2 Real-world Experiment

(a) Environment 1 (b) Environment 2 (c) Environment 3 (d) Environment 4

Figure 4: Real-world experiment and task setup overview. Each of the 4 environments consists of different species of apples (granny smith, golden delicious, red delicious, pink lady, and russet) and a basket. The robot is instructed to "put the two granny smith apples inside a basket". All environments contain exactly two granny smiths while the other species vary, together with presenting different instances of the same species. The agent uses grounding DINO (Liu et al., 2023) to recognise and locate apples and the basket, but cannot accurately distinguish Granny Smiths from other types of apples. Note that using just "granny smith" grounding DINO fails to detect any object in all environments, suggesting limited concept awareness. To evaluate the proposed framework, we conduct experiments with all possible sequences of environments (24 in total) in which the agent starts unaware of concept `grannysmith` and through interaction with the user (both questions and corrections) their belief state is updated. For evaluation we record final cR and $\mathrm{mF1}$ given in table 1.

Agent	cR	cC	mF1%
π_{correct}	-3.70 ± 0.53	-7.70 ± 0.53	50.54 ± 3.72
π_{simple}	-3.45 ± 0.72	-7.45 ± 0.72	45.11 ± 3.08
π_{secure}	**0.47 ± 0.44**†	**-3.625 ± 0.44**†	**51.63 ± 3.99**

Table 1: Real-world experiment results with 95% confidence interval. † indicates a statistically significant difference between agents.

We conduct a real-world experiment on apple manipulation (see figure 4 for overview and appendix C.2 for details). To detect apples, prior weights are assigned using grounding DINO (Liu et al., 2023): when given the prompt "granny smith apple", this localises all the apples and outputs the similarity score between the prompt and localised environment region. This score is used as a prior weight for `grannysmith`. Similarly to the simulation experiments, the robot can attempt to solve the task 5 times before the end of the episode; the user monitors the execution and corrects if a mistake is made. Table 1 shows the summary results. The policy π_{secure} has on average highest cR, cC and $\mathrm{mF1}$. We observe these behaviours:

- π_{secure}: is highly uncertain about the denotations of `grannysmith` and keeps asking for evidence using the question "show me the two granny smiths", after which no reward is obtained for the task (-1 for the question and +1 for task success). After several such questions, the agent can correctly identify denotations and is confident enough to decide to attempt to solve the task;
- π_{simple}: is highly uncertain about the domain and often first opts to ask a question "show me a granny smith", which always results in several corrections on falsely picked fruits afterwards. Importantly, from the evidence gathered, π_{simple} is unable to significantly change its grounding based on the evidence gathered from its previous tasks and always requires some kind of interaction with the user to solve the task.
- π_{correct}: is attempting to solve the task from corrections, leading to the lowest cR, on average requiring 3 attempts to solve the task. At the same time, this gathers a lot of negative evidence, so after attempting to solve several tasks, this policy succeeds later in identifying referents as indicated on average higher $\mathrm{mF1}$ than for π_{simple}.

For additional intuitions about agent behaviours, see appendix C.2 for interaction traces for different agents in table 5 and videos in supplementary material illustrating belief updates using different interaction strategies.

4 Related Work

Lifelong learning SECURE as a framework for lifelong robots is designed for ITL scenarios, where there is natural interaction (embodied conversation in our case) between the agent and the user (Laird et al., 2017). By design, it is not at odds with generative multimodal models (Li et al., 2023; Driess et al., 2023), but rather provides the means of structure-level domain adaptation when embodied agents are deployed in settings in which new concepts are frequently introduced or changed, requiring learning to adapt rapidly by using all possible incidental supervision signals. (Roth, 2017). SECURE handles the issue of unawareness, specifically missing concepts from the hypothesis space of possible domain models, which must be discovered and learned about to solve the task. A related concept is catastrophic forgetting (McCloskey & Cohen, 1989; Ratcliff, 1990), which describes the loss of previously acquired knowledge due to limitations on the agent's memory (i.e., the amount of experience stored). Both of these phenomena hinder lifelong learning but in different ways—e.g., unawareness stems from a lack of knowledge, while forgetting results from the loss of knowledge. SECURE focused on the underexplored problem of unawareness, while catastrophic forgetting was not addressed (our agent had no limit on the number of exemplars it could store).

Language-conditioned Manipulation Language can be used to specify manipulation goals (Stepputtis et al., 2020; Shridhar et al., 2021; Ichter et al., 2022) with grounded representations (Jiang et al., 2023; Gkanatsios et al., 2023), and can be highly modular and generalizable. Our framework enhances such work by providing a methodology to address unawareness. SECURE is most similar to Ren et al. (2023), which explores how planners can be interactively repaired when predictions are confident but wrong. We go beyond this by: a) considering a complete embodied conversation that includes corrections and b) explicitly learning a dialogue policy rather than performing statistical tests (Angelopoulos & Bates, 2021).

Learning from Human Feedback Human-centric agents have to align their behaviour with human preferences (Christiano et al., 2017; Ziegler et al., 2019; Bai et al., 2022) by integrating learning and acting during the fine-tuning phase or continuously through online, human-in-the-loop interactions (Mosqueira-Rey et al., 2023). Traditional models primarily rely on sparse binary rewards, but SECURE leverages symbolic inference to utilise more nuanced embodied conversation messages, enabling it to learn and efficiently perform new tasks involving unforeseen concepts.

Neuro-symbolic Reasoning There is a broad interest in combining neural networks with symbolic/classical algorithms (Garnelo & Shanahan, 2019; Sarker et al., 2021) with many ways to do so (Sheth et al., 2023). In this project, we integrated symbolic reasoning into neural network inference by leveraging semantic analysis of embodied conversation, which builds upon Rubavicius & Lascarides (2022). In the experiments, we study agents that use different semantic analyses to quantify and qualify the benefits of semantic awareness. Large language models have shown the capacity to handle at least some aspects of semantic analysis (Madusanka et al., 2023), but quantifying that capacity is beyond the scope and focus of this work.

5 Conclusion

We have presented a framework for processing embodied conversation to solve rearrangement under unawareness, which broadly falls into the type of task expected to be handled in ITL. We have demonstrated the benefits of making the agent semantics-aware in interactive symbol grounding and task learning, as well as having a mechanism to address false beliefs. There are several directions of future work to consider, including handling and exploiting other semantics of conversation (e.g. generics or contrast) as well as using additional learning signals and knowledge sources.

6 Limitations

This project assumes collaborative human-robot interaction in the context of solving rearrangements that require pick-and-place manipulation of rigid objects. Human-robot interaction may not always be collaborative, and in real-world scenarios, it often involves deception (Shim & Arkin, 2013), which conflicts with the reliance on the user's feedback. The semantic analysis of corrections is explicitly encoded for the set of execution actions (section 2.3.3), so in the case of additional execution actions (e.g., navigation action in the room), additional knowledge needs to be encoded or inferred following semantic analysis of corrections (Lascarides & Asher, 2009). Due to the object-centric view of the environment, SECURE inference complexity (in particular WMC) increases with the number of objects of interest in the environment and the length of the embodied conversation. This, nevertheless, was not an issue in practice due to the limited number of objects and short dialogues. Nonetheless, scaling this to real-world scenarios would require approximate inference (Dubray et al., 2024). Our experiments do not consider dynamic situations, and dialogues were limited to non-fluent properties of objects. In effect, the environment is treated as static, even though its fluent properties (e.g., the location of objects) can change when the agent interacts with the environment. Finally, we are only considering user feedback as a source for learning. Other learning signals, such as utilising data available on the Internet, could also potentially be used to bootstrap the interactive task learning process.

ACKNOWLEDGMENTS

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

M. Appelgren and A. Lascarides. Interactive task learning via embodied corrective feedback. *Journal of Autonomous Agents and Multi-Agent Systems*, 34(54), 2020. doi: doi.org/10.1007/s10458-020-09481-8.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and S. Sabanovic. *Human-Robot Interaction: An Introduction*. Cambridge University Press, 02 2020. ISBN 9781108735407. doi: 10.1017/9781108676649.

Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied AI. *CoRR*, abs/2011.01975, 2020. URL https://arxiv.org/abs/2011.01975.

David Ian Beaver. Chapter 17 - presupposition. In Johan van Benthem and Alice ter Meulen (eds.), *Handbook of Logic and Language*, pp. 939–1008. North-Holland, Amsterdam, 1997. ISBN 978-0-444-81714-3. doi: https://doi.org/10.1016/B978-044481714-3/50022-9. URL https://www.sciencedirect.com/science/article/pii/B9780444817143500229.

Patrick Blackburn and Johan Bos. *Representation and Inference for Natural Language - a First Course in Computational Semantics*. CSLI Studies in Computational Linguistics. CSLI Publications, 2005. ISBN 978-1-57586-496-9. URL http://www.stanford.edu/group/cslipublications/cslipublications/site/1575864967.shtml.

Alexander Bochman. Nonmonotonic reasoning. In Dov M. Gabbay and John Woods (eds.), *The Many Valued and Nonmonotonic Turn in Logic*, volume 8 of *Handbook of the History of Logic*, pp. 557–632. North-Holland, 2007. doi: https://doi.org/10.1016/S1874-5857(07)80012-4. URL https://www.sciencedirect.com/science/article/pii/S1874585707800124.

Norman M Bradburn and Carrie Miles. Vague quantifiers. *Public Opinion Quarterly*, 43(1):92–101, 1979.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

José Miguel Cano Santín, Simon Dobnik, and Mehdi Ghanimifard. Fast visual grounding in interaction: bringing few-shot learning with neural networks to an interactive robot. In Christine Howes, Stergios Chatzikyriakidis, Adam Ek, and Vidya Somashekarappa (eds.), *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pp. 53–61, Gothenburg, June 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.pam-1.7.

Justine Cassell. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Mag.*, 22(4): 67–84, 2001. doi: 10.1609/aimag.v22i4.1593. URL https://doi.org/10.1609/aimag.v22i4.1593.

Mark Chavira and Adnan Darwiche. On probabilistic inference by weighted model counting. *Artif. Intell.*, 172(6-7): 772–799, 2008. doi: 10.1016/j.artint.2007.11.002. URL https://doi.org/10.1016/j.artint.2007.11.002.

Sachin Chitta. Moveit!: an introduction. *Robot Operating System (ROS) The Complete Reference (Volume 1)*, pp. 3–27, 2016.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

Alexandre Dubray, Pierre Schaus, and Siegfried Nijssen. Anytime Weighted Model Counting With Approximation Guarantees For Probabilistic Inference. In *30th International Conference on Principles and Practice of Constraint Programming (CP 2024)*, 2024.

L. T. F. Gamut. *Logic, language, and meaning. Vol.1, Introduction to logic ; L.T.F. Gamut.* University of Chicago Press, Chicago, Ill. ;, 1991. ISBN 0226280853.

Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2018.12.010. URL https://www.sciencedirect.com/science/article/pii/S2352154618301943. Artificial Intelligence.

Nikolaos Gkanatsios, Ayush Jain, Zhou Xian, Yunchu Zhang, Christopher G. Atkeson, and Katerina Fragkiadaki. Energy-based models are zero-shot planners for compositional scene rearrangement. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.030. URL https://doi.org/10.15607/RSS.2023.XIX.030.

H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics Volume 3: Speech Acts*, pp. 41–58. Academic Press, 1975.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski (eds.), *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2022. URL https://proceedings.mlr.press/v205/ichter23a.html.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *Fortieth International Conference on Machine Learning*, 2023.

Oussama Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE J. Robotics Autom.*, 3(1):43–53, 1987. doi: 10.1109/JRA.1987.1087068. URL https://doi.org/10.1109/JRA.1987.1087068.

Saul Kripke. *Naming and Necessity*. Harward University Press, 1980.

John E. Laird, Kevin A. Gluck, John R. Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario D. Salvucci, Matthias Scheutz, Andrea Thomaz, J. Gregory Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21, 2017. doi: 10.1109/MIS.2017.3121552. URL https://doi.org/10.1109/MIS.2017.3121552.

Alex Lascarides and Nicholas Asher. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26 (2):109–158, 2009.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023. doi: 10.48550/arXiv.2301.12597. URL https://doi.org/10.48550/arXiv.2301.12597.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023. doi: 10.48550/ARXIV.2303.05499. URL https://doi.org/10.48550/arXiv.2303.05499.

Tharindu Madusanka, Iqra Zahid, Hao Li, Ian Pratt-Hartmann, and Riza Batista-Navarro. Not all quantifiers are equal: Probing transformer-based language models' understanding of generalised quantifiers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8680–8692, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.536. URL https://aclanthology.org/2023.emnlp-main.536/.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4):3005–3054, 2023. doi: 10.1007/S10462-022-10246-W. URL https://doi.org/10.1007/s10462-022-10246-w.

Tim A. Wheeler Mykel J. Kochenderfer and Kyle H. Wray. *Algorithms for Decision Making*. The MIT Press, Boston Massachusetts, 2022.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. doi: 10.48550/ARXIV.2304.07193. URL https://doi.org/10.48550/arXiv.2304.07193.

Jonghyuk Park, Alex Lascarides, and Subramanian Ramamoorthy. Interactive acquisition of fine-grained visual concepts by exploiting semantics of generic characterizations in discourse. In Maxime Amblard and Ellen Breitholtz (eds.), *Proceedings of the 15th International Conference on Computational Semantics*, pp. 318–331, Nancy, France, June 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.iwcs-1.33.

Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97(2):285–308, 1990. doi: 10.1037/0033-295x.97.2.285.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *CoRR*, abs/2307.01928, 2023. doi: 10.48550/arXiv.2307.01928. URL https://doi.org/10.48550/arXiv.2307.01928.

Dan Roth. Incidental supervision: Moving beyond supervised learning. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 4885–4890. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14950.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950, 2023. doi: 10.48550/ARXIV.2308.12950. URL https://doi.org/10.48550/arXiv.2308.12950.

Rimvydas Rubavicius and Alex Lascarides. Interactive symbol grounding with complex referential expressions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4863–4874, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.358. URL https://aclanthology.org/2022.naacl-main.358.

G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical Report TR 166, Cambridge University Engineering Department, Cambridge, England, 1994.

Bertrand Russell. Knowledge by acquaintance and knowledge by description. In *Mysticism and Logic*, pp. 152–167. London: Longmans Green, 1917.

Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence. *AI Commun.*, 34(3):197–209, jan 2021. ISSN 0921-7126. doi: 10.3233/AIC-210084. URL https://doi.org/10.3233/AIC-210084.

Ami N. Sheth, Kaushik Roy, and Manas Gaur. Neurosymbolic ai - why, what, and how. *ArXiv*, abs/2305.00813, 2023. URL https://api.semanticscholar.org/CorpusID:258426912.

Jaeeun Shim and Ronald C. Arkin. A taxonomy of robot deception and its benefits in hri. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2328–2335, 2013. doi: 10.1109/SMC.2013.398.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13139–13150. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9909794d52985cbc5d95c26e31125d1a-Paper.pdf.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 978-0-262-19398-6. URL https://www.worldcat.org/oclc/37293240.

Alfred Tarski. Sur les ensembles définissables de nombres réels. *Fundamenta Mathematicae*, 17:210–239, 1931. URL https://api.semanticscholar.org/CorpusID:117788493.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005. ISBN 0262201623.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *CoRR*, abs/2404.04125, 2024. doi: 10.48550/ARXIV.2404.04125. URL https://doi.org/10.48550/arXiv.2404.04125.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge J. Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 8927–8948, 2022. doi: 10.1109/TPAMI.2021.3126648. URL https://doi.org/10.1109/TPAMI.2021.3126648.

Zhuo Yang, Yufei Han, Guoxian Yu, and Xiangliang Zhang. Prototypical networks for multi-label learning. *CoRR*, abs/1911.07203, 2019. URL http://arxiv.org/abs/1911.07203.

Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In Jens Kober, Fabio Ramos, and Claire J. Tomlin (eds.), *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, volume 155 of *Proceedings of Machine Learning Research*, pp. 726–747. PMLR, 2020. URL https://proceedings.mlr.press/v155/zeng21a.html.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A LOGIC OF NATURAL LANGUAGE EXPRESSIONS

This appendix outlines the background on representing and reasoning with natural language expressions using logic.

A.1 LOGIC OF SENTENCES

Natural language sentences like "every cube is a cuboid" can be represented in first-order logic. The syntax of the language of first-order logic is constructed recursively: predicate symbols followed by the appropriate number of terms (i.e., a variable $\mathtt{x} \in \mathbb{L}_{\text{var}}$ or a constant $\mathtt{o} \in \mathbb{L}_{const}$) are *well-formed formulae*, and these combine with boolean operators and connectives $\mathbb{L}_{conn} = \{\wedge, \vee, \neg, \rightarrow, \leftrightarrow\}$ and the quantifiers $\forall$ and $\exists$ in the usual way to create more complex well-formed formulae. First-order languages have been used extensively to capture natural language semantics (Blackburn & Bos, 2005): for example, "every cube is a cuboid" has the logical form $\forall\, \mathtt{x}.\mathtt{cube}(\mathtt{x}) \rightarrow \mathtt{cuboid}(\mathtt{x})$.

SECURE utilises generalised quantifier theory to represent the meanings of natural language determiners such as "the one" and "both" (?).[4] generalised quantifier theory extends the classical first-order language with quantifiers additional to $\forall$ and $\exists$, and it is designed to express the meanings of other quantitative sentences about the domain with the usage of quantifiers like e.g., *at least two*, *all but one*, *the one*, *both*, and so on.

Syntactically, a well-formed formula involving a generalised quantifier $\mathtt{Q} \in \mathbb{L}_{quant}$ is of the form $\mathtt{Q}\, \mathtt{x}.(\phi, \psi)$, where $\mathtt{Q}$ binds the variable $\mathtt{x} \in \mathbb{L}_{\text{var}}$, and the restrictor ϕ and body ψ are both well-formed formulae in which $\mathtt{x}$ is absent or free. Very roughly, the restrictor ϕ expresses the descriptive content of the noun phrase that introduced the determiner whose meaning is $\mathtt{Q}$, and the body ψ expresses the content of 'the rest' of the sentence (typically the verb phrase). For example, the natural language sentence "the one red cube is o_1", or more explicitly "the one red cube is object o_1", has the following logical form $\mathtt{\_the\_one\_q}\, \mathtt{x}.(\mathtt{red}(\mathtt{x}) \wedge \mathtt{cube}(\mathtt{x}), \mathtt{object}(\mathtt{o_1}))$. The list of possible generalised quantifiers is given in Table 2 column 1.

Formally, the set of syntactically well-formed formulae—the language of sentences $\mathbb{L}_{snt}$—is defined recursively:

1. If $\mathtt{p}$ is an n-place predicate symbol and $\mathtt{t_1}, \mathtt{t_2}, ..., \mathtt{t_n}$ are terms, then $\mathtt{p}(\mathtt{t_1}, \mathtt{t_2}, ..., \mathtt{t_n})$ is well-formed formula;
2. If ϕ is well-formed formula, then $\neg\phi$ is a well-formed formula;
3. If ϕ and ψ are well-formed formulas, then $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \rightarrow \psi$, $\phi \leftrightarrow \psi$ are all well-formed formulas;
4. If ϕ and ψ are well-formed formulas, known as *restrictor* and *body*, and ϕ and ψ either do not feature the variable $\mathtt{x}$ or $\mathtt{x}$ is free in these formulae, and if $\mathtt{Q}$ is a quantifier, then $\mathtt{Q}\, \mathtt{x}.(\phi, \psi)$, is well-formed formula. For this kind of well-formed formula, the variable $\mathtt{x}$ is a *bound variable* that is bound by $\mathtt{Q}$.

For convenience, we also assume that each object $o \in \mathbb{O}$ in the domain has a unique constant $\mathtt{o}$ that denotes it.

Elements of $\mathbb{L}_{\text{snt}}$ that are well-formed expressions in generalised quantifier theory are assigned a truth-conditional semantics with respect to a first-order domain model $\mathcal{M}$ consisting of a set of objects $\mathbb{O}$, vocabulary $\mathbb{V}$, and the interpretation function $I\colon \mathbb{V} \mapsto \mathbb{O}^*$ as well as a variable assignment function g, which maps variables to individuals in the model (Tarski, 1931). The truth-conditional semantics define a valuation function $[\![\cdot]\!]^{\mathcal{M},g} : \mathbb{L}_{\text{snt}} \mapsto \{0, 1\}$, which in turn specifies when a domain model variable assignment pair $(\mathcal{M}, g)$ satisfies a formula $\phi \in \mathbb{L}_{\text{snt}}$, written $\mathcal{M}, g \models \phi$:

$$\mathcal{M}, g \models \phi \text{ if and only if } [\![\phi]\!]^{\mathcal{M},g} = 1$$

If there are no free variables in $\phi \in \mathbb{L}_{\text{snt}}$ then $[\![\phi]\!]^{\mathcal{M},g} = [\![\phi]\!]^{\mathcal{M},g'}$ for all g, g' and without a loss of generality such situation can be expressed as:

$$\mathcal{M} \models \phi \text{ if and only if } \forall g\ [\![\phi]\!]^{\mathcal{M},g} = 1$$

To define valuation function $[\![\cdot]\!]^{\mathcal{M},g}$, λ-expressions of the form $\lambda\, \mathtt{x}.\phi$ are used in which $\mathtt{x}$ is a variable that is free or absent from ϕ. Note that such λ-expressions are well-formed expressions but are not in $\mathbb{L}_{\text{snt}}$. Here, they are simply used for defining the truth conditions of formulae that feature generalised quantifiers.

[4] SECURE does not consider vague quantifiers (Bradburn & Miles, 1979) like "some" or "few", which implicate soft constraints on objects in the domain model, as a consequence of pragmatic principles of cooperative conversation (Grice, 1975). The interpretation of these quantifiers in cooperative conversation is outside SECURE capabilities but, in principle, could be captured by the agent following the same architecture.

Quantifier $\texttt{Q}$	Surface form	Truth condition $\mathcal{Q}$	Referent constructor $\langle \texttt{Q} \rangle^{\mathcal{M}}$
`_exactly_n_q`	exactly n	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = n\}$
`_at_most_n_q`	at most n	$\lvert \mathbb{R} \cap \mathbb{B} \rvert \leq n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert \leq n\}$
`_at_least_n_q`	at least n	$\lvert \mathbb{R} \cap \mathbb{B} \rvert \geq n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert \geq n\}$
`_a_q`	a/an	$\lvert \mathbb{R} \cap \mathbb{B} \rvert \neq n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert \leq 1\}$
`_every_q`	all/every	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = \lvert \mathbb{R} \rvert \wedge \lvert \mathbb{R} \rvert = 1$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = \lvert \mathbb{O} \rvert \wedge \lvert \mathbb{O} \rvert = 1\}$
`_the_n_q`	the n	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = n \wedge \lvert \mathbb{R} \rvert = n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = \lvert \mathbb{O} \rvert \wedge \lvert \mathbb{O} \rvert = n\}$
`_both_q`	both	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = 2 \wedge \lvert \mathbb{R} \rvert = 2$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = \lvert \mathbb{O} \rvert \wedge \lvert \mathbb{O} \rvert = 2\}$
`_all_but_n_q`	all but n	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = \lvert \mathbb{R} \rvert - n \wedge \lvert \mathbb{R} \rvert > n$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = \lvert \mathbb{O} \rvert - n \wedge \lvert \mathbb{O} \rvert > n\}$
`_n_of_the_m_q`	n of the m	$\lvert \mathbb{R} \cap \mathbb{B} \rvert = n \wedge \lvert \mathbb{R} \rvert = m$	$\{\mathbb{A} \subseteq \mathbb{O} \mid \lvert \mathbb{A} \rvert = n \wedge \lvert \mathbb{O} \rvert = m\}$

Table 2: generalised quantifiers (column 1), surface form (column 2), truth-conditions $\mathcal{Q}$ between the restrictor set $\mathbb{R}$ and body set $\mathbb{B}$, used to evaluate well-formed formulas of the form $\texttt{Q}\,\texttt{x}.(\phi, \psi)$ (column 3), and referent constructor used to evaluate the logical form of referential expressions for the domain model $\mathcal{M}$ (column 4).

Well-formed formulas in truth-conditional semantics are either evaluated as being true (1) or false (0). This evaluation is given recursively:

$$\llbracket \texttt{a} \rrbracket^{\mathcal{M},g} = \begin{cases} I(\texttt{a}) & \text{if a is a predicate symbol} \\ g(\texttt{a}) & \text{if a is a variable} \\ I(\texttt{a}) & \text{if a is a constant}^{5} \end{cases}$$

$$\begin{aligned}
\llbracket \texttt{p}(\texttt{t}_1, \texttt{t}_2, \ldots, \texttt{t}_\texttt{n}) \rrbracket^{\mathcal{M},g} &= 1 \text{ if } (\llbracket \texttt{t}_1 \rrbracket^{\mathcal{M},g}, \llbracket \texttt{t}_2 \rrbracket^{\mathcal{M},g}, \ldots, \llbracket \texttt{t}_n \rrbracket^{\mathcal{M},g}) \in \llbracket \texttt{p} \rrbracket^{\mathcal{M},g} \\
\llbracket \neg\phi \rrbracket^{\mathcal{M},g} &= 1 \text{ if } \llbracket \phi \rrbracket^{\mathcal{M},g} = 0 \\
\llbracket \phi \wedge \psi \rrbracket^{\mathcal{M},g} &= 1 \text{ if } \llbracket \phi \rrbracket^{\mathcal{M},g} = 1 \text{ and } \llbracket \psi \rrbracket^{\mathcal{M},g} = 1 \\
\llbracket \phi \vee \psi \rrbracket^{\mathcal{M},g} &= 1 \text{ if } \llbracket \phi \rrbracket^{\mathcal{M},g} = 1 \text{ or } \llbracket \psi \rrbracket^{\mathcal{M},g} = 1 \\
\llbracket \phi \rightarrow \psi \rrbracket^{\mathcal{M},g} &= 1 \text{ if } \llbracket \phi \rrbracket^{\mathcal{M},g} = 0 \text{ or } \llbracket \psi \rrbracket^{\mathcal{M},g} = 1 \\
\llbracket \phi \leftrightarrow \psi \rrbracket^{\mathcal{M},g} &= 1 \text{ if } \llbracket \phi \rightarrow \psi \rrbracket^{\mathcal{M},g} = 1 \text{ and } \llbracket \psi \rightarrow \phi \rrbracket^{\mathcal{M},g} = 1 \\
\llbracket \lambda\,\texttt{x}.\phi \rrbracket^{\mathcal{M},g} &= \{o \in \mathbb{O} \mid \llbracket \phi \rrbracket^{\mathcal{M},g[\texttt{x}/\texttt{o}]} = 1\} \\
\llbracket \texttt{Q}\,\texttt{x}.(\phi, \psi) \rrbracket^{\mathcal{M},g} &= \mathcal{Q}(\llbracket \lambda\,\texttt{x}.\phi \rrbracket^{\mathcal{M},g}, \llbracket \lambda\,\texttt{x}.\psi \rrbracket^{\mathcal{M},g})
\end{aligned}$$

where $g[\texttt{x}/\texttt{o}]$ is a variable assignment function just like g except that $g[\texttt{x}/\texttt{o}](\texttt{x}) = \texttt{o}$ and $\mathcal{Q}$ is a truth-condition (Table 2 column 3) used to evaluate well-formed formulas that feature generalised quantifiers. The value of these truth conditions depends on two sets defined by the λ-expression, commonly referred to as the restrictor set $\mathbb{R}$ and the body set $\mathbb{B}$.

A.2 Logic of Referential Expressions

Natural language expressions that denote objects in the environment are known as referential expressions r (Kripke, 1980). Linguistically, they are expressed by noun phrases: a noun phrase can consist of a determiner and a noun with pre- and post-modifiers (e.g., "the one red cube that is to the left of every green sphere"); a pronoun or demonstrative (e.g., "it", "this", "that"), or proper name (e.g., "Edinburgh", "Kim", " Mr. Smith"). SECURE focuses on referential expressions that are of the first kind: i.e.; they include a determiner and a noun with pre-and/or post-modifiers. We denote the logical form of referential expression r as $\Phi(r)$ and the language of these well-formed expressions as $\mathbb{L}_{\text{ref}}$.

Since r is a noun phrase and not a full sentence like the ones discussed in section A.1, $\Phi(r)$ consists of a generalised quantifier $\texttt{Q}$ (corresponding to r's determiner), its restrictor ϕ (corresponding to r's adjectives and nouns) but the body of the quantifier is 'missing'. In semantic analysis, this is traditionally captured using λ-calculus with logical forms like $\lambda\,\psi.\texttt{Q}\,\texttt{x}.(\phi, \psi)$ but this is not what we want to capture to process embodied conversation. The λ-term denotes the set of properties satisfied by the objects that satisfy ϕ, with quantifier $\texttt{Q}$ imposing constraints on the relationship between the denotations of ϕ and ψ (and sometimes also on ϕ itself as well). In our scenario, however, we need

[5] As mentioned earlier, we assume a unique constant denotes each object, and so without loss of generality, we can assume $I(\texttt{a}) = a$, where $a \in \mathbb{O} \wedge \mathbb{O} \in \mathcal{M}$ is the unique object denoted by $\texttt{a}$.

to identify the set of referent(s) that are denoted by the referring expression r, not the properties satisfied by those referents. Because of these requirements, we express the logical form $\Phi(r)$ of the referential expression r another way and interpret it using the theory of generalised quantifiers (**?**). Specifically, we represent the referential expression "the one red cube" with the following logical form $\langle \texttt{\_the\_one\_q}\ \texttt{x}.\texttt{red}(\texttt{x}) \wedge \texttt{cube}(\texttt{x})\rangle$. More generally, the logical forms of referential expressions are of the form $\langle \texttt{Q}\ \texttt{x}.\phi\rangle$, where $\phi \in \mathbb{L}_{\mathrm{snt}}$ is a well-formed formula with only one free variable $\texttt{x}$. Intuitively, this formula captures the content of the described objects and $\texttt{Q}$ captures the content of the determiner. The notation $\langle\cdot\rangle$ is a convenient way to differentiate these logical forms from well-formed formulae and in particular, the logical forms of sentences.

The logical form of a referential expression $\Phi(r) \in \mathbb{L}_{\mathrm{ref}}$ is evaluated for the domain model $\mathcal{M}$ to yield the referent $\mathcal{R}$. An element of $\mathcal{R}$ is a *set* of objects because r may denote more than one object (e.g., any denotation of "two cubes" consists of two objects). $\mathcal{R}$ itself is a set of (potentially) more than one element because $\mathcal{M}$ may have more than one set of objects that are denoted by the referring expression r: e.g., if there is more than one object in $\mathcal{M}$ that is a cube, then "a cube" has more than one referent in $\mathcal{M}$. For example, given the domain model $\mathcal{M}$ with objects $\mathbb{O} = \{o_1, o_2\}$ and interpretation function $I(\texttt{cube}) = \mathbb{O}$, the expression "all cubes" should denote $\{\{o_1, o_2\}\}$ (there is only one set denoted by "every cube", and that is the set of all objects in the domain that are cubes), while "a cube" denotes $\{\{o_1\}, \{o_2\}\}$ (i.e., there are two sets of objects that are denoted by this referential expression). Thus more formally, $\mathcal{R}$ is always a subset of the power set of $\mathbb{O}$: $\mathcal{R} \subseteq 2^{\mathbb{O}}$. These examples reveal two factors that affect the referent of $\Phi(r)$. First, objects in a denotation of $\Phi(r)$ satisfy the restrictor of the generalised quantifier.[6] Secondly, the generalised quantifier imposes its conditions on $\mathcal{R}$, in particular on the cardinality of each set in $\mathcal{R}$ that is a denotation, and for some quantifiers there is also a constraint on the relationship between a denotation and all the objects $\mathbb{O}$, which is equivalent to a constraint on the cardinality of $\mathcal{R}$ itself. For instance, each set in a referent for "at least two cubes" must have a cardinality of at least two; for "exactly two cubes" the cardinality must be equal to 2; and "the two cubes" and "both" impose the additional constraint that not only should each denotation have cardinality 2, but also this denotation is *unique* (Russell, 1917) (i.e., $\mathcal{R}$ itself has cardinality 1). Further, the referent for "every cube" is unique because it is the (unique) maximal set of objects in $\mathcal{M}$ that satisfy the restrictor (again, $\mathcal{R}$ must have cardinality 1). Constraints imposed by different quantifiers are lead to logical consequences about the environment (Beaver, 1997). Note that such constraints are not English-specific: Finnish "molempi" has the same constraint as English "both". With this in mind, we obtain a formal definition of the reference semantics for an arbitrary logical form of the referential expression $\Phi(r) = \langle \texttt{Q}\ \texttt{x}.\phi\rangle$ by first defining a projection of $\mathcal{M}$ onto a smaller domain model $\sigma(\mathcal{M}, \phi, \texttt{x})$, which consists of *all and only* those objects $o \in \mathbb{O}$ that satisfy $\phi[\texttt{x}/\texttt{o}]$ (i.e., the formula ϕ with each occurrence of $\texttt{x}$ substituted with the unique constant $\texttt{o}$ that denotes $o \in \mathbb{O}$):

$$\begin{aligned} \sigma(\mathcal{M}, \phi, \texttt{x}) = \sigma((\mathbb{O}, \mathbb{V}, I), \phi, \texttt{x}) &= (\mathbb{O}', \mathbb{V}, I') \\ \mathbb{O}' &= \{o \in \mathbb{O} \mid \mathcal{M} \models \phi[\texttt{x}/\texttt{o}]\} \\ I' &= I \downarrow \mathbb{O}' \text{ (i.e., } I \text{ projected onto } \mathbb{O}') \end{aligned} \tag{9}$$

We then define the reference semantics of $\Phi(r)$ for this projected domain model. As discussed before, the truth-conditional semantics of generalised quantifiers are defined using the restrictor $\mathbb{R}$ and body $\mathbb{B}$. But in our scenario, consisting only of referring expressions, the body $\mathbb{B}$ is not present for truth-conditional evaluation. However, by evaluating the content of the quantifier for the smaller projected model and making $\mathbb{B} = \mathbb{O}'$ (i.e., the objects in that projected model), the reference semantics can focus solely on how they constrain the *cardinalities* of these sets to achieve the above-desired effects for constructing the referent $\mathcal{R}$ Table 2 column 4 shows how the referent is defined using set constructors. These constraints and the definition of model projection in Eq. 9 yield the reference semantics (Eq. A.2) for the logical form $\Phi(r)$ with the desired properties:

$$\mathcal{R} = \Phi(r)^{\mathcal{M}} = \langle \texttt{Q}\ \texttt{x}.\phi\rangle^{\mathcal{M}} = \langle \texttt{Q}\rangle^{\sigma(\mathcal{M}, \phi, \texttt{x})}$$

where $\langle \texttt{Q}\rangle^{\mathcal{M}}$ is a referent constructor, utilising the condition specific to the (generalised) quantifier (and defined in Table 2 column 4).

[6] This work ignore group nouns like "committee"

B Algorithmic Details

Algorithm 1 outlines the belief state update procedure using a well-formed formula, constructed from the semantic analysis of embodied conversation and algorithm 2 outlines dialogue strategy optimization procedure using a semi-gradient SARSA with hyperparameter values given in table 3.

Algorithm 1 Update (Belief state update with well-formed formula)

Require: Belief state $b \in \mathbb{B}$ consisting of a set of objects $\mathbb{O}$ set of object-centric embeddings $\mathbb{X}$ (one-to-one corresponce with objects), vocabulary of predicates $\mathbb{V}$, support of embedding-label vector pairs $\mathbb{S}$, domain theory build throughout the embodied conversation Δ, prior weights w_p, and grounded weights w_g as well as a well-formed formula $\phi \in \mathbb{L}_{\text{snt}}$.

$\mathbb{S}' \leftarrow \emptyset$ ▷ Initialize new support
$\Delta \leftarrow \Delta \cup \{\phi\}$ ▷ Update domain theory
for $o, \boldsymbol{x}$ in $\text{Zip}(\mathbb{O}, \mathbb{X})$ **do** ▷ Iterate over corresponding object-embedding pairs
    $\mathtt{o} \leftarrow \text{Const}(o)$ ▷ Get constant of an object $o \in \mathbb{O}$
    $y \leftarrow \text{Dictionary}()$ ▷ Dictionary for symbol predictions
    for $\mathtt{p}$ in $\mathbb{V}$ **do** ▷ Iterate over predicate symbols
        $y[\mathtt{p}] \leftarrow \text{CON}_{w_p}(\mathtt{p}(\mathtt{o}) \mid b)$ ▷ Conditional probability with prior weights for an atom $\mathtt{p}(\mathtt{o})$
    $\boldsymbol{y} \leftarrow \text{Vectorize}(y)$ ▷ Construct semantic vector of size $|\mathbb{V}|$
    $\mathbb{S}' \leftarrow \mathbb{S}' \cup (\boldsymbol{x}, \boldsymbol{y})$ ▷ Add exemplar to new support
$\mathbb{S} \leftarrow \mathbb{S}'$ ▷ Update support
for $o, \boldsymbol{x}$ in $\text{Zip}(\mathbb{O}, \mathbb{X})$ **do** ▷ Iterate over corresponding object-embedding pairs
    for $\mathtt{p}$ in $\mathbb{V}$ **do** ▷ Iterate over predicate symbols
        $w_g(\mathtt{p}(\mathtt{o})) \leftarrow \omega_b(\boldsymbol{x})^{(\mathtt{p})}$ ▷ Grounding model prediction
return b

Algorithm 2 Dialogue Strategy Optimisation using Episodic Semi-gradient SARSA

Require: belief state $b \in \mathbb{B}$, state-action value function Q parameterised by θ, number of tasks per environment m, learning rate α, discount factor γ, epsilon ϵ.

while not converged **do**
    $\mathcal{E} \leftarrow \text{GetEnvironment}()$
    for t in $\text{GetTasks}(\mathcal{E}, m)$ **do** ▷ Get m tasks in environment $\mathcal{E}$
        $b \leftarrow \text{Update}(b, \Phi(t))$ ▷ Update belief state (Algorithm 1)
        $\mathbb{A}_{\text{quest}} \leftarrow \text{GetQuestions}(t)$ ▷ Coherent questions for t
        $\mathbb{A} \leftarrow \{a_{\text{act}}\} \cup \mathbb{A}_{\text{quest}}$ ▷ Construct action space
        while True **do** ▷ Training loop
            $a \leftarrow \arg\max_{a \in \mathbb{A}} Q(b, a)$ ▷ Greedy action
            $R, \phi \leftarrow \text{Act}(\mathcal{E}, a)$ ▷ Act in $\mathcal{E}$ and construct ϕ via semantic analysis
            $b' \leftarrow \text{Update}(b, \phi)$ ▷ Update belief state (Algorithm 1)
            if $|R| = 1$ **then** ▷ Terminating state
                $\delta \leftarrow R - Q(s, a)$
                $\theta \leftarrow \theta + \alpha\delta\nabla_\theta Q(s, a)$
                break
            else
                $a' \leftarrow \begin{cases} \arg\max_{a \in \mathbb{A}} Q(b', a) \text{ with probability } 1 - \epsilon \\ \text{Choose}(\mathbb{A}) \text{ otherwise} \end{cases}$ ▷ ϵ-greedy action
                $\delta \leftarrow R + \gamma Q(b', a') - Q(b, a)$
                $\theta \leftarrow \theta + \alpha\delta\nabla_\theta Q(b, a)$
            $b \leftarrow b'$ ▷ Update belief state

Hyparameter	Symbol	Value
Number of tasks per environment	m	1
Learning rate	α	0.1
Discount factor	γ	0.99
Exploration factor	ϵ	0.1

Table 3: Hyperparameter values used for optimizing dialogue strategies for π_{secure} and π_{simple} with semi-gradient SARSA following algorithm 2.

C Experiment Details

The following appendix provides additional details for simulation and real-world experiments.

C.1 Simulation Experiments

Simulation experiments are conducted in the MuJoCo simulator in rearranging rigid objects.

For obtaining the logical forms of referential expressions that are part of embodied conversation messages (task instructions, questions, corrections) we used a simple semantic parser based on CodeLlama7B (Rozière et al., 2023)[7] with in-context learning and prompt given in figure 5. For the grounding model, we utilize object-centric embeddings by first using ground-truth object locations from the simulator to localize patches in the environment's top-down view (as in figure 2c), from which the feature vectors $\boldsymbol{x} \in \mathbb{R}^{384}$ are obtained using DINOv2 (Oquab et al., 2023)[8] encoder. In inference, $\tau = 0.65$ is set as the threshold for adding examples to the positive/negative support, which results in an example being added to the negative support if $\hat{y}^{(\text{p})} \leq 0.354$ and added to the positive support if $\hat{y}^{(\text{p})} \geq 0.646$.

Agent's dialogue strategy is given by action-value function (Eq.8) which is optimized offline. To learn dialogue strategy parameters $\theta = [\theta_1, \theta_2]^\top$ for π_{secure} and π_{simple} (π_{correct} does not make interaction-level decisions and just uses corrective feedback), we used semi-gradient SARSA (Rummery & Niranjan, 1994) for policy optimization using tasks in simulation whose instructions feature subsets of domain concepts. In particular, we use all quantifiers and spatial relationships elicited in figure 2 but only a subset of properties (plain, cube, red, green, and blue). For the simulation experiments, we set unit costs of symbol designation and reference resolution to be the same and small to encourage exploration when learning new concepts. In particular we set $C_{\text{point}} = C_{\text{ref}} = 0.1$.

When the agent decides to rearrange the environment, given by choosing an action $a_{\text{act}} \in \mathbb{A}$, this manipulation is achieved using an operational space controller Khatib (1987) and a rudimentary logic for setting control targets based on the pose of the object of interest. After each action, the oracle is checked, and in case of the suboptimal action, a correction is issued, and the previous action is undone (e.g. returning the picked object to the original position).

C.2 Real-world Experiment

For real-world experiments, we modify several aspects of the experimental setup compared to simulation experiments while keeping the rest the same. In the real-world, we use the ZED 2i stereo camera (just the left feed) as the visual sensor. Using it, grounding DINO (Liu et al., 2023)[9]: an open-set grounding model is given the prompt "granny smith apples" that returns candidate patches from the environment. We set the threshold for prediction retrieval to 0.3 (so regions that are identified to locate "granny smith apples" below 0.3 are not considered). Each patch has an associated object-centric embedding $\boldsymbol{x} \in \mathbb{R}^{256}$. For the apple manipulation, we set $C_{\text{point}} = 0.2$ and $C_{\text{ref}} = 0.6$ to make querying more expensive and encourage the agent to query only on high uncertainty situations—these arise at the beginning of interaction when a neologism is encountered. When the agent decides to execute the plan, it performs pick-and-place moves using the MoveIt (Chitta, 2016) to plan and execute collision-free trajectories with pick-and-place locations determined using Transporter Networks (Zeng et al., 2020) trained on generic pick-and-place tasks.

Tables 4 and 5 provide additional visualization of the real-world environment, task setup, and qualitative behaviour differences via learning traces.

[7] https://huggingface.co/codellama/CodeLlama-7b-Instruct-hf
[8] https://huggingface.co/facebook/dinov2-small
[9] https://huggingface.co/IDEA-Research/grounding-dino-tiny

Prompt for Semantic parsing.

[INST] You are a helper in a virtual assistant. You are asked to translate referential expressions to their logical forms. Referential Expressions like "the one red block" have a logical form $\langle \texttt{\_the\_1\_q}\ \texttt{x}.\texttt{red}(\texttt{x}) \wedge \texttt{block}(\texttt{x})\rangle$. Logical forms follow the pattern $\langle[\texttt{Quantifier}][\texttt{Variable}].\ [\texttt{Formula}]\rangle$ $[\texttt{Quantifier}]$ is a quantifier. We consider the following quantifiers:

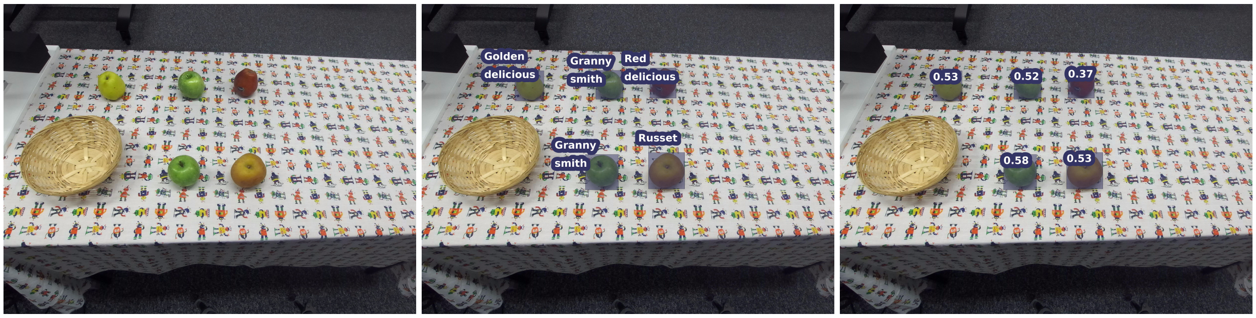
- existential with surface form: "a/an" and denoted as logic symbol: $\texttt{\_a\_q}$
- universal with surface form: "every/all" and denoted as a logic symbol: $\texttt{\_every\_q}$
- uniqueness with surface form: "the n" and denoted as a logic symbol: $\texttt{\_the\_n\_q}$, where n is a natural number like "one", "two", etc.

$[\texttt{Variable}]$ is a variable, e.g. $\texttt{x}$, $\texttt{x1}$, $\texttt{x12}$. Note that $[\texttt{Variable}]$ is the only free variable in $[\texttt{Formula}]$ not bound by a quantifier. if the formula has only one variable, it should be named $\texttt{x}$.
$[\texttt{Formula}]$ is a formula of predicate logic. Each formula is constructed recursively:

- Predicates like $\texttt{red}(\texttt{x})$, $\texttt{above}(\texttt{x1}, \texttt{x2})$, $\texttt{left}(\texttt{x1}, \texttt{x3})$ is a well-formed formula $[\texttt{Formula}]$
- Negation of $[\texttt{Formula}]$ like $\texttt{neg}(\texttt{red}(\texttt{x}))$ is a well-formed formula $[\texttt{Formula}]$
- Logical conjunction of $[\texttt{Formula}]$s like $\texttt{red}(\texttt{x}) \wedge \texttt{block}(\texttt{x})$ is a well-formed formula $[\texttt{Formula}]$
- Logical disjunction of $[\texttt{Formula}]$s like $\texttt{red}(\texttt{x}) \vee \texttt{block}(\texttt{x})$ is a well-formed formula $[\texttt{Formula}]$
- Logical implication of $[\texttt{Formula}]$s like $\texttt{red}(\texttt{x}) \rightarrow \texttt{block}(\texttt{x})$ is a well-formed formula $[\texttt{Formula}]$
- Logical structure for generalised quantifiers $[\texttt{Quantifier}][\texttt{Variable}].([\texttt{Formula}], [\texttt{Formula}])$ like $\texttt{\_the\_1\_qx}.(\texttt{red}(\texttt{x}), \texttt{block}(\texttt{x}))$ is a well-formed formula $[\texttt{Formula}]$

Here are some examples of what these logical forms of referential expressions look like:
RefExp, LF-RefExp,
"a block.", $\texttt{\_a\_q}\ \texttt{x}.\texttt{block}(\texttt{x})\rangle$
"the one block.", $\langle\texttt{\_the\_1\_q}\ \texttt{x}.\texttt{block}(\texttt{x})\rangle$
"the two plain objects.", $\langle\texttt{\_the\_2\_q}\ \texttt{x}.\texttt{plain}(\texttt{x}) \wedge \texttt{object}(\texttt{x})\rangle$
"every magenta sphere." $\langle\texttt{\_every\_q}\ \texttt{x}.\texttt{magenta}(\texttt{x}) \wedge \texttt{sphere}(\texttt{x})\rangle$
"not a block above a sphere.", $\langle\texttt{\_a\_q}\ \texttt{x}.\texttt{neg}(\texttt{\_a\_q}\ \texttt{x1}.(\texttt{sphere}(\texttt{x1}), \texttt{block}(\texttt{x}) \wedge \texttt{above}(\texttt{x}, \texttt{x1})))\rangle$
"a sphere to the left of every green cone.",
$\langle\texttt{\_a\_q}\ \texttt{x}.\texttt{\_every\_q}\ \texttt{x1}.(\texttt{green}(\texttt{x1}) \wedge \texttt{cone}(\texttt{x1}), \texttt{sphere}(\texttt{x}) \wedge \texttt{left}(\texttt{x}, \texttt{x1}))\rangle$,
"every sphere to the left of every green object.",
$\langle\texttt{\_every\_q}\ \texttt{x}.\texttt{\_every\_q}\ \texttt{x1}.(\texttt{green}(\texttt{x1}) \wedge \texttt{object}(\texttt{x1}), \texttt{sphere}(\texttt{x}) \wedge \texttt{left}(\texttt{x}, \texttt{x1}))\rangle$,
"a sphere to the right of the two green cones.",
$\langle\texttt{\_a\_q}\ \texttt{x}.\texttt{\_the\_2\_q}\ \texttt{x1}.(\texttt{green}(\texttt{x1}) \wedge \texttt{cone}(\texttt{x1}), \texttt{sphere}(\texttt{x}) \wedge \texttt{right}(\texttt{x}, \texttt{x1}))\rangle$
"the one sphere in front of every green cone.",
$< \texttt{\_the\_1\_q}\ \texttt{x}.\texttt{\_every\_q}\ \texttt{x1}.(\texttt{green}(\texttt{x1}) \wedge \texttt{cone}(\texttt{x1}), \texttt{sphere}(\texttt{x}) \wedge \texttt{front}(\texttt{x}, \texttt{x1}))\rangle$
"the two spheres behind a green cone.",
$\langle\texttt{\_the\_2\_q}\ \texttt{x}.\texttt{\_a\_q}\ \texttt{x1}.(\texttt{green}(\texttt{x1}) \wedge \texttt{cone}(\texttt{x1}), \texttt{sphere}(\texttt{x}) \wedge \texttt{behind}(\texttt{x}, \texttt{x1}))\rangle$
Now, please translate the following referential expression to its logical form. Just give the logical form. No extra information. Referential expression: {refexp} [/INST]

Figure 5: Prompt for parsing referential expressions to their logical forms. {refexp denotes the referential expression to be parsed using this prompt.



(a) Environment 1

(b) Environment 2

(c) Environment 3

(d) Environment 4

Table 4: Visualisation of apple manipulation environments, their ground-truth labels, and initial grounding DINO predictions *before* interaction with the environment and the user via embodied conversation.

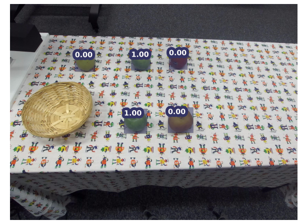
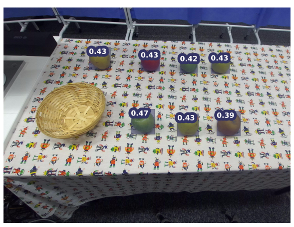
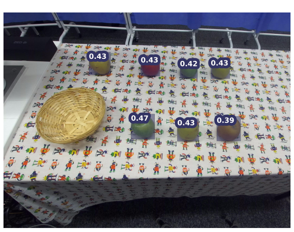
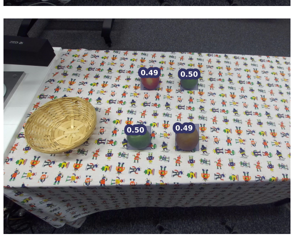
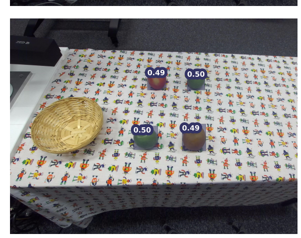
<table>
<tr><th>Initial belief state</th><th>Embodied conversation</th><th>Updated belief state</th></tr>
<tr><td colspan="3">π_{correct}</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Okay. Let's try to achieve the task
User: No. This is a golden delicious.
Agent. Sorry. Let's try again.
User: No. This is a russet.
Agent. Sorry. Let's try again.
User: No. This is a red delicious.
Agent. Sorry. Let's try again.
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Okay. Let's try to achieve the task
User: No. This is a red delicious.
Agent. Sorry. Let's try again.
User: No. This is a golden delicious.
Agent. Sorry. Let's try again.
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
<tr><td colspan="3">π_{simple}</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Before that, show me a granny smith.
User: (designates one of the granny smiths)
Agent: Okay. Let's try to achieve the task
User: No. This is a golden delicious.
Agent. Sorry. Let's try again.
User: No. This is a golden delicious.
Agent. Sorry. Let's try again.
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Okay. Let's try to achieve the task
User: No. This is a russet.
Agent. Sorry. Let's try again.
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
<tr><td colspan="3">π_{secure}</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Before that, show me the two granny smiths.
User: (designates the two granny smiths)
Agent: Okay. Let's try to achieve the task
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
<tr><td>
</td><td>User: Put the two granny smiths inside the basket.
Agent: Okay. Let's try to achieve the task
Agent: I have finished the task.
User: Correct.</td><td>
</td></tr>
</table>

Table 5: Interaction traces for real-world experiments showing different agent behaviour in two consequent environments when attempting to solve the task instructed. Each trace shows the initial, and updated belief state as well as the embodied conversation the user and agent engage in.