ASK & EXPLORE: GROUNDED QUESTION ANSWERING FOR CURIOSITY-DRIVEN EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many real-world scenarios where extrinsic rewards to the agent are extremely sparse, curiosity has emerged as a useful concept providing intrinsic rewards that enable the agent to explore its environment and acquire information to achieve its goals. Despite their strong performance on many sparse-reward tasks, existing curiosity approaches rely on an overly holistic view of state transitions, and do not allow for a structured understanding of specific aspects of the environment. In this paper, we formulate curiosity based on grounded question answering by encouraging the agent to ask questions about the environment and be curious when the answers to these questions change. We show that natural language questions encourage the agent to uncover specific knowledge about their environment such as the physical properties of objects as well as their spatial relationships with other objects, which serve as valuable curiosity rewards to solve sparse-reward tasks more efficiently.

1 INTRODUCTION

Efficient exploration in the absence of dense reward signals is a long-standing problem in reinforcement learning (Vecerik et al., 2017). Without dense extrinsic signals, a promising alternative is to define suitable auxiliary *intrinsic* signals that can help the agent in exploring its environment (Laud, 2004). Recently, *curiosity* has emerged as a promising computational framework for modeling intrinsic reward and has brought major advances in many sparse-reward domains (Pathak et al., 2017; Burda et al., 2018a;b; Pathak et al., 2019; Dean et al., 2020). While the algorithmic details of different methods vary, the core idea is to use changes in the observed state as the intrinsic reward to encourage agents to explore their environment. Despite their strong performance on many sparse-reward tasks, these existing approaches tend to rely on a holistic view of state transitions and do not allow for a targeted understanding of specific aspects of the environment. However, not all states are equally interesting but such information is not available to the agent a priori. On the contrary, humans rely on extensive knowledge about the world when exploring the environment. Language serves as a powerful medium for encoding this knowledge. A particular type of language that humans use is *question* – in an unfamiliar environment, humans often start the exploration by asking what can be done in the environment. Based on this observation, we hypothesize that *language-based question answering* may provide a grounded and targeted medium to probe specific knowledge about the current state in order to solve the task at hand.

As a step towards more structured and flexible curiosity-driven learning, we develop a novel form of curiosity, ASK & EXPLORE (ANE), that leverages *grounded question answering* to encourage the agent to ask questions about the environment and be curious when the answers to these questions change. These questions can capture physical properties of the objects (e.g., *Is the large sphere green in color?*) as well as their spatial relationships with other objects (e.g., *Are there any blue spheres behind the cyan ball?*). By using language as a compositional medium to uncover specific knowledge about the environment, we are able to train an agent that explores and solves challenging long-horizon sparse-reward tasks. In addition to our qualitative results, we perform an in-depth study of what type of questions are useful under what scenarios to provide empirical guidelines for applying our method.

2 BACKGROUND

Exploration bonuses and curiosity-driven exploration. Exploration bonuses motivate agents to explore their environment even when extrinsic reward r_t^e is sparse (or zero) by training the policy to maximize a new reward $r_t = r_t^e + r_t^i$, where r_t^i is the exploration bonus or the intrinsic reward at time t (Krebs et al., 2009; Dayan & Sejnowski, 1996; Sutton, 1990). The intrinsic reward r_t^i is

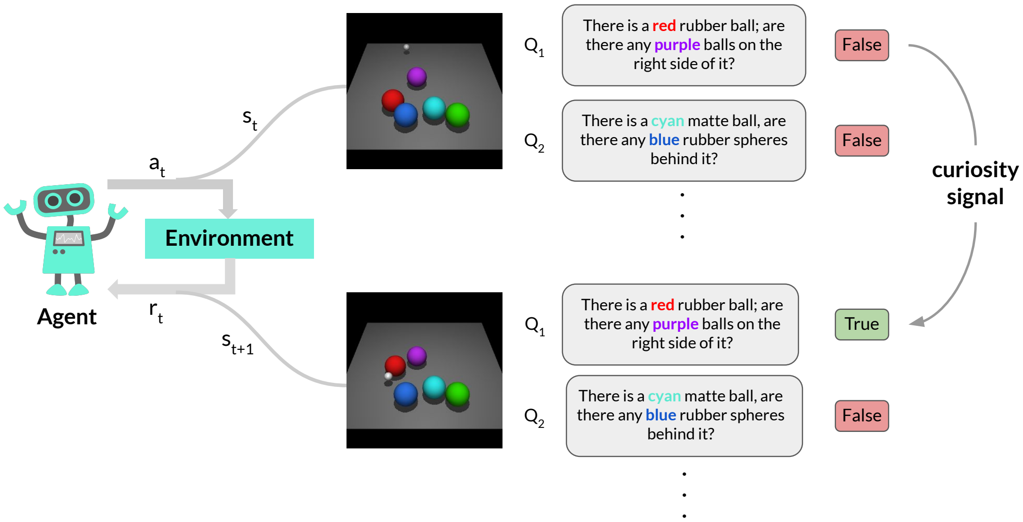

Agent
Environment
a_t
r_t
s_t
s_{t+1}
Q_1
There is a red rubber ball; are there any purple balls on the right side of it?
False
Q_2
There is a cyan matte ball, are there any blue rubber spheres behind it?
False
curiosity signal
Q_1
There is a red rubber ball; are there any purple balls on the right side of it?
True
Q_2
There is a cyan matte ball, are there any blue rubber spheres behind it?
False

Figure 1: **Ask & Explore**: Our approach proposes a curiosity formulation that leverages grounded question answering to query specific knowledge about the environment. The agent is encouraged to ask questions and be curious about transitions when the answer to a question changes (details in Section 3).

designed to be higher in novel states in order to encourage the agents to explore less frequently visited states. In recent years, several promising algorithms in this family include: 1) Curiosity-driven exploration by self-supervised prediction (Pathak et al., 2017; Burda et al., 2018a; Pathak et al., 2019), which formulates an intrinsic reward that encourages the agent to favor transitions with high prediction error using dynamics-based learning, and 2) Random Network Distillation (RND) (Burda et al., 2018b), which encourages novelty by training the policy to minimize the prediction error of a predictor neural network as it tries to mimic a randomly initialized target neural network. Several other approaches have used count-based exploration (Bellemare et al., 2016; Tang et al., 2017) and multimodal signals (Dean et al., 2020) to encourage exploration.

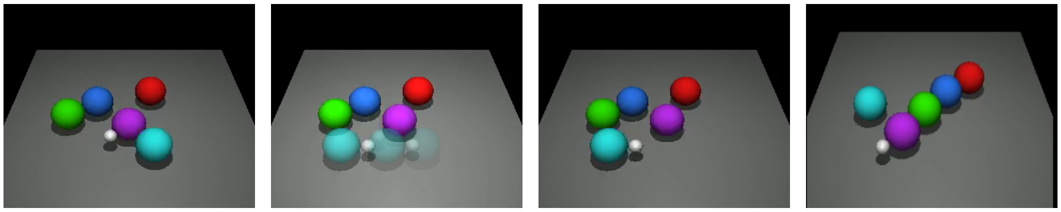
CLEVR-Robot Environment: We perform experiments using the CLEVR-Robot Environment (Jiang et al., 2019), an open-source object interaction environment built using the MuJoCo physics engine (Todorov et al., 2012) and CLEVR language engine (Johnson et al., 2017a). The environment is designed to serve as a testbed for studying grounded language understanding and object manipulation. To succeed in this environment, the agent must be able to handle a varying number of objects with diverse visual and physical properties (see details in Appendix A).

3 Ask & Explore

Our goal is to develop an intrinsic reward that leverages the knowledge about the physical properties of objects and how objects in the environment relate to each other. Such intrinsic reward may bridge the gap between passive pattern recognition and active decision making. To design the intrinsic reward, we chose grounded language as a flexible medium for encoding this knowledge. The CLEVR-Robot environment provides an ideal testbed for using grounded language as a source of intrinsic reward as it provides functionalities to generate scenes and language (in the form of questions) that can be evaluated as the agent interacts with the environment. Note that access to the true state and language is not required. Indeed, for an agent in the real world, such assumptions do not hold. Nonetheless, just like humans can describe a scene with language, the agent can also be equipped with a parameterized visuolinguistic model such as a visual question answering model (VQA) or image captioning model. We plan to explore these directions in future works.

We formulate an intrinsic reward that aims to generate the agent's curiosity about transitions when the answers to the agent's questions grounded in the environment change. At every step, the agent has access to n questions, $Q_1, Q_2, ...Q_n$. For question Q_k, the difference in the answer before the transition $Q_k(s_t)$, and after the transition $Q_k(s_{t+1})$ contributes to the curiosity signal corresponding to that action (Figure 1). To evaluate $Q_k(s_t)$, we experiment with two types of intrinsic rewards that leverage language. One of them uses the labeling function of the CLEVR-Robot environment and the other utilizes a parameterized VQA model (experiments with the latter can be found in Appendix D.2). The intrinsic reward at time t, r_t^i, expressed as

$$r_t^i = \sum_{k=1}^{n} \mathbb{1}[Q_k(s_t) \neq Q_k(s_{t+1})] \tag{1}$$



(a) Goal is: "*There is a* ***green*** *sphere; are there any rubber* ***cyan*** *balls* ***in front of*** *it?*".

(b) Agent performs actions and interacts with the environment and tries to satisfy goal.

(c) Resulting state: Agent receives **+1** reward.

(d) Object ordering

Figure 2: We use the CLEVR-Robot environment and consider both dense (a-c) and sparse (d) reward settings. The global location of the objects vary across episodes (images from Jiang et al. (2019)).

Further algorithmic details such as how the questions are selected can be found in Appendix B.

4 EXPERIMENTS

We design our experiments to understand the following overarching question: *Does an agent with grounded language understanding explore the environment in a more structured and efficient manner?* To answer different aspects of this question, we first evaluate our approach in tasks with different reward sparsity (Section 4.1). Then, we evaluate the impact of different types of grounded language understanding on the performance and how the impact differs in settings with varying reward sparsity (Section 4.2). Finally, we study the effect of the linguistic feedback's density on the efficacy of exploration (due to space limit, we defer the details to Appendix D.1).

We compare our approach to three baselines:

1. Proximal Policy Optimization (PPO) (Schulman et al., 2017) (no exploration bonus)

2. Intrinsic Curiosity Module (ICM) (Pathak et al., 2017)

3. Random Network Distillation (RND) (Burda et al., 2018b).

We use the same optimized hyperparameters from the original papers (Pathak et al., 2017; Burda et al., 2018b). The agent is trained using PPO in all experiments with the same hyperparameters[1]. We perform three independent runs of each algorithm without any tuning of random seeds, and plot the mean and standard deviation across the three runs (see details in Appendix C).

4.1 VARYING DEGREE OF REWARD SPARSITY

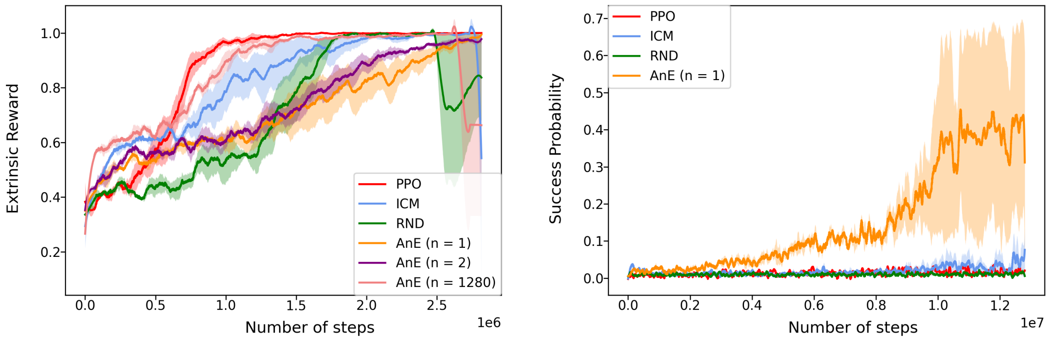
To study in what scenarios grounded language understanding can help exploration, we test our approach in two tasks with drastically different reward sparsity.

Dense reward setting. In this setting, the agent needs to complete an object alignment goal where the spatial relationship between two objects in the environment is specified, for example, *"There is a green sphere; are there any rubber cyan balls in front of it?"* (Figure 2 (a-c)). The agent receives a reward of +1 if it manipulates the objects to achieve the desired spatial arrangement. When the environment is reset before every episode, it is ensured that the goal state is not satisfied initially.

Sparse reward setting. In the **object ordering** task (Figure 2 (d)), the agent needs to order the objects by color in a single line, for example, *"arrange the objects so that their colors range from blue to green in the horizontal direction, and keep the objects close vertically"*. The ordering of colors we specify is: *cyan, purple, green, blue, red* from left to right. The agent is given a +10 reward if it is able to successfully order the objects in this arrangement, and 0 otherwise.

The results for the two settings are shown in Figure 3. We observe in the dense reward setting, PPO with no exploration bonus outperforms all curiosity-driven methods, which highlights that curiosity does not provide a significant advantage when the reward is dense. In the sparse reward task, we find that while all existing baseline methods struggle to make meaningful progress, ANE significantly outperforms the baselines in this setting using a single question (n = 1) at each step. This confirms our hypothesis – an intrinsic reward that leverages grounded language understanding is better at exploring the environment. The exploration results in a wider coverage of relevant states and helps the agent learn to solve the task more efficiently compared to existing novelty-based exploration

[1] The original ICM uses A3C but we used PPO similar to Burda et al. (2018a).



(a) Dense reward setting

(b) Sparse reward setting

Figure 3: AnE significantly outperforms the baselines (PPO, ICM and RND) in the sparse reward setting which demonstrates the effectiveness of an intrinsic reward based on grounded question answering.

methods. In addition, we study the impact of scaling to multiple questions in both dense and sparse reward environments in Appendix D.1, and the performance of different approaches in the absence of any extrinsic reward in Appendix D.3.

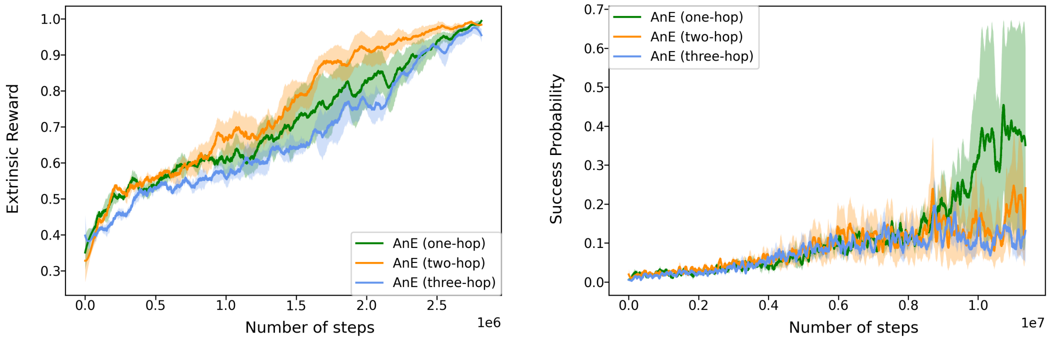
4.2 Varying complexity of questions

To better understand which questions are most useful for the task, we test the performance of AnE using three types of questions querying varying complexity of spatial relationships between objects in the environment - **one**, **two** and **three** "hop" questions:

One-hop: *"There is a red metallic sphere; are there any green matte balls left of it?"*

Two-hop: *"Are there any purple rubber balls that are on the left side of the red sphere that is behind the blue matte ball?"*

Three-hop: *"There is a green rubber ball behind the red metallic sphere; are there any blue balls in front of the purple matte sphere?"*



(a) Dense reward setting

(b) Sparse reward setting

Figure 4: Comparing the performance of AnE for questions of varying complexity.

We observe in Figure 4 how the performance of the agent varies with increasing complexity of questions (in terms of a greater number of pair-wise relationships between objects). It is interesting to note that the relationship between language and agent performance is not the same across different task settings. The extrinsic reward increases in the dense reward setting as we move from one-hop to two-hop questions in the environment and then decreases as we progress to three-hop questions. While in the sparse task, the success rate is highest using one-hop questions and decreases as we increase the number of object relationships, although all improve over the baselines. Therefore, even simple probes of spatial relationships are sufficient as a curiosity signal.

5 Conclusion

In this paper, we proposed a novel form of curiosity by encouraging the agent to ask questions about the environment and be curious when the answers to their questions change. We show that this formulation of intrinsic reward probes targeted knowledge about the physical properties of the objects as well as their spatial relationships with other objects, achieving significantly better performance than existing curiosity methods on highly sparse reward tasks.

REFERENCES

Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2018a.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018b.

Peter Dayan and Terrence J Sejnowski. Exploration bonuses and dual control. *Machine Learning*, 25(1):5–22, 1996.

Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. *arXiv preprint arXiv:2007.03669*, 2020.

Yiding Jiang, Shixiang Gu, Kevin Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. *arXiv preprint arXiv:1906.07343*, 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017a.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2989–2998, 2017b.

Ruth M Krebs, Björn H Schott, Hartmut Schütze, and Emrah Düzel. The novelty exploration bonus and its attentional modulation. *Neuropsychologia*, 47(11):2272–2281, 2009.

Adam Daniel Laud. Theory and application of reward shaping in reinforcement learning. Technical report, 2004.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pp. 5062–5071. PMLR, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pp. 216–224. Elsevier, 1990.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *31st Conference on Neural Information Processing Systems (NIPS)*, volume 30, pp. 1–18, 2017.

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*, pp. 5026–5033. IEEE, 2012.

Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018.

APPENDIX

A CLEVR-ROBOT ENVIRONMENT

The CLEVR-Robot environment (Jiang et al., 2019) was designed in MuJoCo for object manipulation tasks. In the environment, the agent can interact with objects with diverse visual and physical properties. The environment supports CLEVR style (Johnson et al., 2017a) language so the agent may receive linguistic feedback as it interacts with the environment.

We use the *discrete* action space which consists of a point mass agent pushing 1 of 5 objects in 1 of the 8 cardinal directions for a fixed number of frames, so the discrete action space has size 40. Please refer to (Jiang et al., 2019) for details of the action space. As a proof of concept for using grounded language for exploration, we consider the standard 5 objects setting which contains a fixed set of 5 spheres of different colors- *cyan, purple, green, blue, red.* We plan to experiment with the diverse objects setting where objects can take up different shapes such as *cube, sphere* and *cylinder* in future work.

The environment supports all CLEVR style language. For the experiments, we consider 3 types of language statement: *one-hop*, *two-hop* and *three-hop*. The number indicates the number of objects involved in the spatial reasoning (for an *h-hop question*, *h+1* objects are involved) and, indirectly, the complexity of reasoning. The number of hops also affects the density of intrinsic reward and what kind of states the agent is encouraged to visit.

B IMPLEMENTATION DETAILS

We begin with a fixed set of questions S from which a subset of n questions, $Q_1, Q_2, ...Q_n$ is sampled at each step of the episode. A counter is maintained for a large reservoir of possible questions to record the frequency of answer flips corresponding to each question. We use a hyperparameter $0.5 \leq \alpha \leq 1$ as a threshold to set an upper bound on the % of answer flips any question can encounter when it is sampled, after which it is replaced by a new question sampled from the reservoir (e.g., If α = 0.6 and Q_1 has witnessed 650 answer flips out of the 1000 times it was sampled, it is replaced by a new question Q_k which has not been seen by the agent yet). This is an attempt to ensure that if the agent has learned transitions to exploit a particular language statement, it does not continue to exploit it.

Algorithm 1 provides a more complete picture of the approach.

C EXPERIMENTAL DETAILS

We use one copy of the environment since CLEVR-Robot Environment does not support multithreading currently. We used rollouts of length 128 in all experiments. We use 3 optimization epochs per rollout for our approach and ICM, whereas 4 epochs for RND. The episode terminates either if the agent achieves the goal or exceeds maximum time steps. The agent is provided a sparse terminal binary reward only if it arranges the objects according to the spatial relationship defined by the goal (e.g., arrange the objects horizontally according to some color ordering), and 0 otherwise.

Table 1 contains details of how we preprocessed the environment for our experiments.

Hyperparameter	Value
Grey-scaling	False
Observation downsampling	(64,64)
Extrinsic reward clipping	False
Intrinsic reward clipping	False

Table 1: Preprocessing details for the environments for all experiments.

We refer to the following open-source repositories for baselines:

ICM: https://github.com/pathak22/noreward-rl
RND: https://github.com/openai/random-network-distillation
ICM (Pytorch implementation): https://github.com/jcwleo/curiosity-driven-exploration-pytorch
RND (Pytorch implementation): https://github.com/jcwleo/random-network-distillation-pytorch

Algorithm 1 ANE pseudo-code

```
$N$ ← number of rollouts
$N_{\text{opt}}$ ← number of optimization steps
$K$ ← length of rollout
$S$ ← set of questions initialized
$C$ ← counter for questions in $S$ initialized to 0
$n$ ← number of questions queried at each step
$D$ ← initialized with $n * K$ questions from $S$
$\alpha$ ← threshold which determines maximum answer flipping frequency for a question
Sample state $s_0 \sim p_0(s_0)$
for $i = 1$ to $K$ do
    for $j = 1$ to $n$ do
        sample $q \sim S$
        add $q$ to $D[i]$
        remove $q$ from $S$
    end for
end for
$t = 0$
for $\beta = 1$ to $N$ do
    intrinsic reward $r_t^i = 0$
    shuffle entries in $D$
    for $j = 1$ to $K$ do
        $Q_1, Q_2, ...Q_n = D[j]$
        evaluate $Q_1(s_t), Q_2(s_t), ...Q_n(s_t)$
        sample $a_t \sim \pi(a_t \mid s_t)$
        sample $s_{t+1}, r_t^e \sim p(s_{t+1}, r_t^e \mid s_t, a_t)$
        evaluate $Q_1(s_{t+1}), Q_2(s_{t+1}), ...Q_n(s_{t+1})$
        for $k = 1$ to $n$ do
            if $Q_k(s_t) \neq Q_k(s_{t+1})$ then
                $r_t^i$ += 1
                $C[Q_k]$ += 1
                if $C[Q_k] \;/\; \beta \geq \alpha$ then
                    replace $Q_k$ with new question $q$ from $S$ (question at index 0)
                    remove $q$ from $S$
                end if
            end if
        end for
        add $s_t, s_{t+1}, a_t, r_t^e, r_t^i$ to optimization batch $B_\beta$
        t += 1
    end for
    Calculate target $T_\beta$ and advantage $A_\beta$
    for $j = 1$ to $N_{\text{opt}}$ do
        optimize $\theta_\pi$ wrt PPO loss on batch $B_\beta, T_\beta, A_\beta$ using Adam
    end for
end for
```

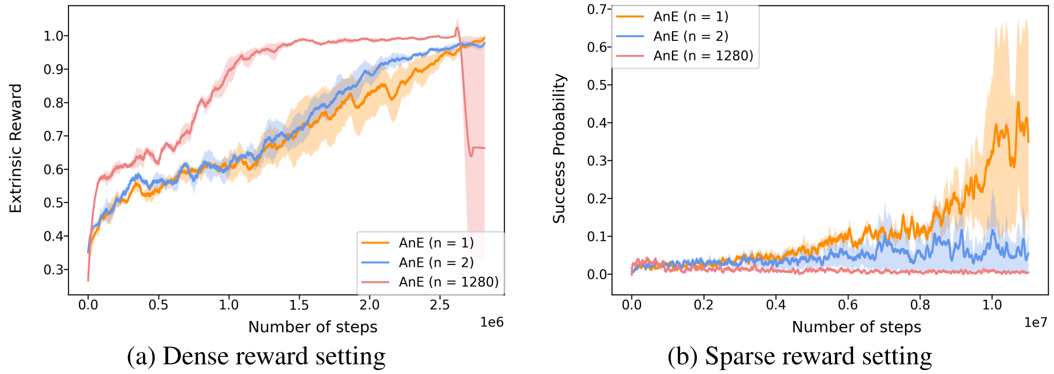
D ADDITIONAL RESULTS

We present additional analysis on several design decisions in our approach.

D.1 VARYING NUMBER OF QUESTIONS

We study the impact of number of questions used by the agent to query the environment on its performance in both dense and sparse reward settings. The results are shown in Figure 5.

It is interesting to observe that effect of linguistic feedback's density changes across different settings. For object alignment tasks in the dense reward setting, the agent's performance has an increasing relationship with the number of questions (n) asked at each step. On the contrary, in the sparse reward settings, we notice that increasing the number of questions does not help the agent explore better, and the agent's curiosity is declining with increasing n. We believe this difference can be



(a) Dense reward setting (b) Sparse reward setting

Figure 5: Comparing the performance of ANE for different values of n.

attributed to the highly different nature of the tasks in terms of both complexity and also the potential inherent impossibility of simultaneously achieving high intrinsic reward and extrinsic reward – this effect is magnified when the reward is extremely sparse (in the ordering task), so a trade-off is needed to be made in terms of linguistic feedback density. It would be interesting to see if it is possible to automatically find such balance.

D.2 VQA Model for curiosity-driven exploration

We demonstrate that an agent equipped with a parameterized VQA model possesses grounded language understanding and hence can leverage our approach for curiosity-driven exploration (Figure 6). We train a CNN-LSTM model used as a baseline in Johnson et al. (2017b). We plan to work with more sophisticated and interpretable approaches in future works which represent human language as programs (Johnson et al., 2017b; Yi et al., 2018) to scale up to diverse object settings and eventually to settings where ground truth language is not available such as navigation.

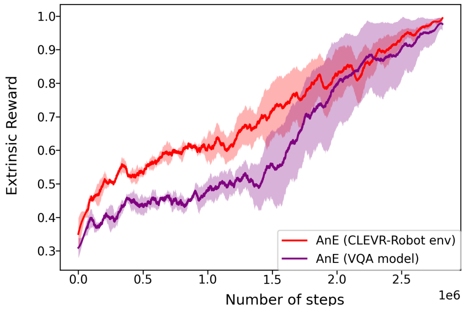


Figure 6: Performance of ANE using a VQA model for grounded question answering

D.3 Pure Exploration

We compare the performance of ANE against curiosity-based baseline methods ICM and RND using pure exploration agents (agent does not have access to extrinsic reward) (Figure 7). We observe that even in the absence of extrinsic reward ANE performs comparable to ICM and better than RND in the dense reward setting. In the sparse reward settings, all methods were unable to achieve success solely using the intrinsic reward, which suggests that the effect of ANE comes from more than having denser extrinsic reward signal.

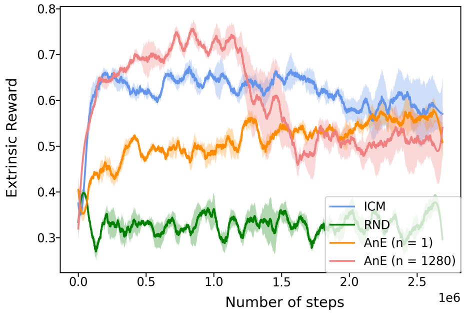


Figure 7: Comparing the performance of ANE against baselines ICM and RND in the absence of any extrinsic reward