See, Hear, Explore: Curiosity via Audio-Visual Association

Victoria Dean
Carnegie Mellon University
vdean@cs.cmu.edu

Shubham Tulsiani
Facebook AI Research
shubhtuls@fb.com

Abhinav Gupta
Carnegie Mellon University
Facebook AI Research
abhinavg@cs.cmu.edu

Abstract

Exploration is one of the core challenges in reinforcement learning. A common formulation of curiosity-driven exploration uses the difference between the real future and the future predicted by a learned model [1]. However, predicting the future is an inherently difficult task which can be ill-posed in the face of stochasticity. In this paper, we introduce an alternative form of curiosity that rewards novel associations between different senses. Our approach exploits multiple modalities to provide a stronger signal for more efficient exploration. Our method is inspired by the fact that, for humans, both sight and sound play a critical role in exploration. We present results on several Atari environments and Habitat (a photorealistic navigation simulator), showing the benefits of using an audio-visual association model for intrinsically guiding learning agents in the absence of external rewards. For videos and code, see https://vdean.github.io/audio-curiosity.html.

1 Introduction

Many successes in reinforcement learning (RL) have come from agents maximizing a provided extrinsic reward such as a game score. However, in real-world settings, reward functions are hard to formulate and require significant human engineering. On the other hand, humans explore the world driven by intrinsic motivation, such as curiosity, often in the absence of rewards. But what is curiosity and how would one formulate it?

Recent work in RL [1–3] has focused on curiosity using future prediction. In this formulation, an exploration policy receives rewards for actions that lead to differences between the real future and the future predicted by a forward dynamics model. In turn, the dynamics model improves as it learns from novel states. While the core idea behind this curiosity formulation is simple, putting it into practice is quite challenging. Learning and modeling forward dynamics is still an open research problem; it is unclear how to handle multiple possible futures, whether to explicitly incorporate physics, or even what the right prediction space is (pixel space or some latent space).
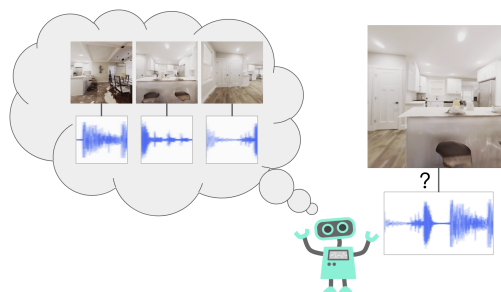

Figure 1: **See, Hear, Explore**: We propose a formulation of curiosity that encourages the agent to explore novel associations between modalities, such as audio and vision. In Habitat, shown above, our method allows for more efficient exploration than baselines.

The use of multiple modalities in *human* learning has a long history. Research in psychology has suggested that humans look for incongruity [4]. A baby might hit an object to hear what it sounds like. Have you ever found yourself curious to touch a material different from anything you have seen before? Humans are drawn towards discovering and exploring novel associations between different modalities. Dember and Earl [5] argued that intrinsic motivation arises with discrepancy between expected sensory perception and the actual stimulus. More recent work has shown the presence of multimodal stimulation and exploration in infants [6, 7]. In cognitive development, both sight and sound guide exploration: babies are drawn towards colorful toys that squeak and rattle [8].

Inspired by human exploration, we introduce See Hear Explore (SHE): a curiosity for novel associations between sensory modalities (Figure 1). SHE rewards actions that generate novel associations (shared information) between different sensory modalities (in our case, pixels and sounds). We first demonstrate that our formulation is useful in several Atari games: SHE allows for more exploration, is more sample-efficient, and is more robust to noise compared to existing curiosity baselines on these environments. Finally, we show experiments on area exploration in the realistic Habitat simulator [9]. Our results demonstrate that in this setting our approach significantly outperforms baselines.

To summarize, our contributions in this paper include: 1) SHE, a curiosity formulation that searches for novel associations in the world. To the best of our knowledge, multimodal associations have not been investigated in self-supervised exploration; 2) we show our approach outperforms the commonly-used curiosity approaches on standard Atari benchmark tasks; 3) most importantly, multimodality is one of the most basic facets of our rich physical world (audio and vision are generated by the same physical processes [10]). We show experiments on realistic area exploration in which SHE significantly outperforms baselines. This work builds on efficient exploration, which will be crucial as we push agents to explore more complex unknown environments.

2 Related Work

Our work uses audio as an additional modality for self-supervised exploration. We divide the prior work into two categories: exploration (Section 2.1) and multimodal learning (Section 2.2).

2.1 Exploration

Prior work on exploration has used error [11, 12, 1, 13], uncertainty [14, 15, 3], and potential improvement [16] of a prediction model as intrinsic motivation. Some approaches have used count-based or pseudo-count-based exploration [17, 18]. Others use auxiliary losses to supplement reward functions and improve sample efficiency [19, 20].

One popular approach to self-supervised exploration is curiosity by self-supervised prediction [1, 2]. In this form of curiosity, an intrinsic reward encourages an agent to explore situations with high error under a jointly-trained future prediction model. The model's error is a proxy for novelty: unpredictable situations are more likely novel and therefore ones the agent should explore. These future-predicting models can be difficult to train, especially in visual space. Our method also looks at self-supervised exploration, but our intrinsic reward does not rely on future prediction. We circumvent the need for predicting the future by leveraging multimodal input. SHE rewards association classification error (i.e. association novelty) as opposed to higher-dimensional prediction error. Our key insight is that *associative* models across modalities are simpler to learn, and their accuracy is also indicative of novelty.

2.2 Multimodal Learning

Multimodal settings are especially amenable to self-supervision, as information from one modality can be used to supervise learning for another modality. One prior work learned a joint visual and language representation using Flickr images and associated descriptors [21]. In computer vision, audio can provide additional information that complements images [22–24]. Recent work [25, 26] has looked at audio-visual embodied navigation, in which audio is emitted from a goal point to aid in supervised learning of navigation. In the same environment, Gao et al. [27] used audio and visual information for learning visual feature representations. We test on this audio-visual navigation environment, but for unsupervised exploration in RL; we have no goal states.

Audio and visual information are closely linked, and since we commonly have access to both in the form of video, this is a rich area for self-supervision. Aytar et al. [28] used audio from Atari in the form of YouTube videos of people playing the games. This work uses audio-visual demonstrations from YouTube to learn a visual embedding. The setup here is learning from demonstrations from humans. In our case, on the other hand, the audio-visual associations drive intrinsically motivated exploration. We learn multimodal alignment from active data, which the agent both collects and uses.

In robotics settings, the use of additional modalities such as tactile sensing [29, 30] or audio [31] is increasingly popular for grasping and manipulation tasks. Lee et al. [32] showed the effectiveness of self-supervised training of tactile and visual representations by demonstrating its use on a peg insertion task. While these previous approaches have demonstrated the benefits of using multiple sensory modalities for learning better representations or accurately solving tasks, in this work we demonstrate its utility for allowing agents to explore. To the best of our knowledge, using audio to learn actions for exploration is unique to our work.

3 See, Hear, Explore

We now describe SHE, our exploration method based on associating audio and visual information. Our goal is to develop a form of curiosity that exploits the multimodal nature of the input data. Our core idea is that the SHE agent learns a model that captures associations between two modalities. We use this model to reward actions that lead to unseen associations between the modalities. By rewarding such actions, we guide the exploration policy towards discovering new combinations of sight and sound.

More formally, we consider an agent interacting with an environment that contains visual and sound features, which we call $x_t = (v_t, s_t)$ for time t where v_t is the visual feature vector and s_t is the sound feature vector. The agent explores using a policy $a_t \sim \pi(v_t; \theta)$ where a_t corresponds to an action taken by the agent at time t. To make for easier comparison to visual-only baselines, our agent is only given access to the visual features v_t and not the audio features s_t. To enable this agent to explore, we train a discriminator D that tries to determine whether an observed multimodal pair (v_t, s_t) is novel, and we reward the agent in states where the discriminator is surprised by the observed multimodal association.

3.1 Why Novel Associations?

The goal of an exploration policy is to perform actions that uncover states that lead to a better understanding of the world. One commonly used exploration strategy involves rewarding actions that lead to unseen or novel states [17]. While this strategy seems intuitive, it does not handle the fact that while some states might not have been seen, we still understand them and hence they do not need to be explored. In light of this, recent approaches have used a prediction-based formulation. If a model cannot predict the future, it needs more data points to learn. However, sometimes we may have seen enough examples, and prediction is still challenging, leading a prediction-based exploration policy to get stuck. For example, consider the couch-potato issue: the random TV in the Unity environment (as described in Burda et al. [2]) yields high error for prediction models, so prediction-based curious agents receive high rewards for staring at the TV, though this is not a desirable type of exploration.

Trying to avoid these problems has shaped much of the work on intrinsic motivation; Schmidhuber [16], Oudeyer et al. [33], White et al. [34], and Burda et al. [35] all formulate intrinsic rewards with the goal of mitigating problems like the couch-potato agent. Our approach, different from this body of prior work, looks at how multimodal data can mitigate these issues.

Our underlying hypothesis is that discovering new sight and sound associations will help mitigate the shortcomings of the previously described count-based and prediction-based exploration strategies. By using an association model, we ask a simpler question: can this image co-occur with this sound? Consider another example, in which pressing a button randomly produces one among 3 distinct sounds. Our approach could learn to classify all as associated, while an agent using future prediction error would always be curious. This focus on association effectively helps ignore stochasticity, mitigating the couch-potato problem by focusing on non-random structure. Such a model can allow generalization to unseen states, and it also does not need to predict the future to provide an informative signal for exploration.
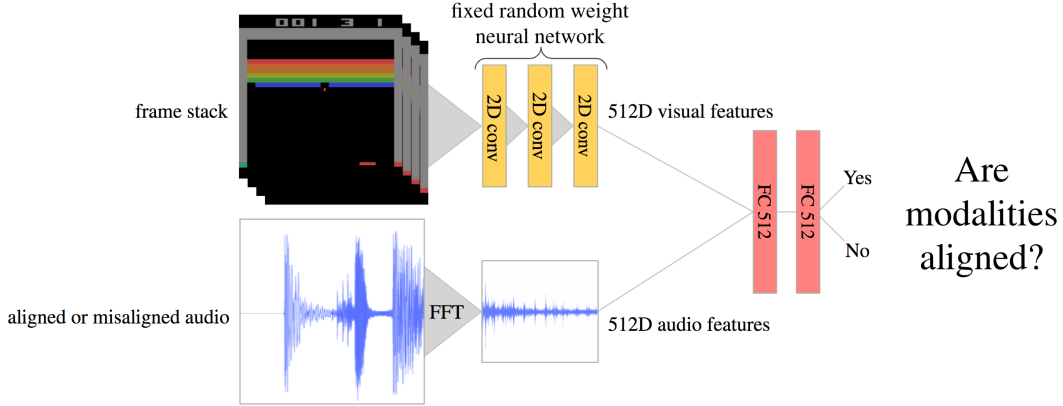
Figure 2: **Our audio-visual association model**: The frames (top left) and potentially misaligned audio waveform (bottom left) are preprocessed into 512-dimensional feature vectors using a random feature network and FFT, respectively. The discriminator network (right) takes these features as inputs and is trained to output whether or not they are aligned. 2D conv represents a standard convolutional layer and FC 512 represents a fully-connected layer with 512 units.

3.2 Association Novelty via Alignment

The core of our method is the ability to determine whether a given pair (v_t, s_t) represents a novel association. To tackle this problem, we learn a model in an online manner. Given past trajectories, a model learns whether a certain audio-visual input comes from a seen or new phenomenon. One way to model this would be to use a generative model such as a VAE [36] or GAN [37], which could determine if the image-audio combination is within the distribution or out of distribution. However, generative models are also difficult to train, so we instead propose using a discriminator to predict if the image-audio pair is novel, which has a much smaller, binary output space.

We train this discriminator to distinguish real audio-visual pairs from 'fake' pairs from another distribution, with the insight that the learned model is more likely to classify novel pairs as fake. Here, the observed image-audio pairs during exploration act as positive training examples, but a critical question is how to obtain negative image-audio pairs. To this end, we reformulate the problem as whether image-audio pairs are aligned or not: we obtain 'fake' samples by randomly misaligning the audio and visual modalities, similar to Owens and Efros [23]. The positive data is then the aligned image-audio pairs, and the negative data is comprised of misaligned ones. The discriminator model, as shown in Figure 2, outputs values between 0 and 1, with 1 representing high probability of audio-visual alignment and 0 representing misalignment. We can then leverage the misalignment likelihood as an indicator of novelty since the discriminator would be uncertain in such instances.

3.3 Training

Having introduced association novelty via alignment, we now describe how we implement this idea using function approximators. During training, the agent policy is rolled out in parallel environments. These yield trajectories which are each chunked into 128 time steps. A trajectory consists of pairs of preprocessed visual and sound features: $(v_1, s_1), (v_2, s_2)...(v_{128}, s_{128})$. These trajectories are used for two purposes: 1) updating the discriminator D as described below and 2) updating the exploration policy based on the intrinsic reward r_t^i (computed using the discriminator), also described below.

Training the Alignment Discriminator The discriminator D is a neural network that takes a visual and sound feature pair as input and outputs an alignment probability. To train D, we start with positive examples from the visual and sound feature pairs (v_t, s_t). With 0.5 probability we use the true aligned pair, and with 0.5 probability we create a false pair consisting of the true visual feature vector v_t and a sound feature vector uniformly sampled from the current trajectory. We call this false sound s_t'. We define a binary variable z_t to indicate whether the true audio was used, i.e. when we give the discriminator the true audio s_t, we set $z_t = 1$, and when we give the discriminator the false audio s_t',

4

$z_t = 0$. We use a cross-entropy loss to train the discriminator, similar to prior work [23, 28]:

$$\mathcal{L}_t(v_t, s_t, z_t) = \begin{cases} -\log(D(v_t, s_t)), & \text{if } z_t = 1 \\ -\dfrac{||s_t - s'_t||_2}{\mathbb{E}_{\text{batch}}||s_t - s'_t||_2} \log(1 - D(v_t, s'_t)), & \text{if } z_t = 0 \end{cases}$$

In the $z_t = 0$ case above, we weight the cross-entropy loss to prevent the discriminator from being penalized in cases where the true and false audio are similar. We weight by the L2 difference between the true and false audio feature vectors and normalize by dividing by the mean difference across samples in the batch of 128 trajectories. This loss is used for updating the discriminator and is not used in computing the agent's intrinsic reward.

Training the Agent via Intrinsic Reward We want to reward actions that lead to unseen image-audio pairs. For a given image-audio pair, if the discriminator predicts 0 (unseen or unaligned), we want to reward the agent. On the other hand, if the discriminator correctly outputs 1 on a true pair, the agent receives no reward. Mathematically, the agent's intrinsic reward is the negative log-likelihood of the discriminator evaluated on the true pairs: $r_t^i := -\log(D(v_t, s_t))$, where the output of D is between 0 and 1. Audio-visual pairs that the discriminator knows to be aligned get a reward of 0, but if the discriminator is uncertain (the association surprised the discriminator) the agent receives a positive reward. The agent takes an action and receives a new observation v_t and intrinsic reward r_t^i (note that the agent does not have access to the sound s_t). The agent is trained using PPO [38] to maximize the expected reward: $\max_\theta \mathbb{E}_{\pi(v_t;\theta)} \left[\sum_t \gamma^t r_t^i \right]$. *The agent does not have access to the extrinsic reward. Extrinsic reward is used only for evaluation.* This will enable the use of our method on future tasks for which we cannot easily obtain a reward function. See Burda et al. [2] for further discussion on training with no extrinsic reward while using it for evaluation.

4   Experiments

In this section, we will test our method in two exploration settings (Atari and Habitat) and compare it with commonly-used curiosity formulations.

4.1   Environments

Atari Similar to prior work, we demonstrate the effectiveness of our approach on 12 Atari games. We chose a subset of the Atari games to represent environments used in prior work and a range of difficulty levels. We excluded some games due to lack of audio (e.g. Amidar, Pong) or the presence of background music (e.g. RoadRunner, Super Mario Bros). The action space is different from the one used in the future prediction curiosity work [2], as we use Gym Retro [39] in order to access game audio, and Retro environments use a larger action space. The original work reported results using the minimal action space, Discrete(4), whereas we use Discrete(6). We note that the larger action space does slow exploration, but it is used for both our method and the baselines for fair comparisons. To compute audio features, we take an audio clip spanning 4 time steps (1/15th of a second for these 60 frame per second environments) and apply a Fast Fourier Transform (FFT). The FFT output is downsampled using max pooling to a 512-dimensional feature vector, which is used as input to the discriminator along with a 512-dimensional visual feature vector.

Habitat Navigation We also test our method in a navigation setting using Habitat [9] (Figure 3). In this environment, the agent moves around a photorealistic Replica scene [40]. We use the largest Replica scene, Apartment 0, which has 211 discrete locations. In each location, the agent can face in 4 directions. At each timestep, the agent takes one of 3 discrete actions: turn left, turn right, or move forward. As in our Atari experiments, the agent is not given any extrinsic reward; we simply want to see how well it can explore the area without supervision. We use the audio-visual navigation extension from Chen et al. [25], which emits a fixed audio clip from a fixed location and allows our agent to hear the sound after simulating room acoustics. The perceived sound at each time step is less than 1 second long, and we zero pad this audio to 1 second to make each sound equal length for feature computation. We apply FFT and downsample to a 512-dimensional feature vector, the same as done in Atari, described above.

Figure 3: **Habitat visualization**: Left: an example agent view. Right: the top-down map for apartment 0 (not seen by agent). The agent is the blue arrow and the audio source is the green square. Gray areas are open space, while white areas are obstacles, which make exploration challenging.

4.2 Baselines

We compare to future prediction curiosity [2], which as previously described performs visual future prediction. We build upon the open-source code from the authors (see the appendix for more details). We also compare to exploration via disagreement [3] and Random Network Distillation (RND) [35]. We use the same hyperparameters (which were optimized for the future prediction and disagreement baselines) for policy learning across all approaches. We use random CNN features [2, 35] for the visual feature representation for our method and the baselines in all experiments.

4.3 Atari Experimental Results

We trained our approach and baselines for 200 million frames using the *intrinsic* reward and measure performance by the *extrinsic* reward throughout learning. Figure 4 shows these results. Each method was run with three random seeds, and the plots show the mean and standard error for each method. Please see the appendix for more experimental details. Across many environments, our method enables better exploration (as judged by the extrinsic reward) and is more sample efficient than the baselines. Of the 12 environments, SHE outperforms the disagreement baseline in 9 and the future prediction baseline in 8. We hypothesize that states leading to novel audio-visual associations, such as a new sound when killing an enemy, are more indicative of a significant event than ones inducing high prediction error (which can happen due to inaccurate modeling or stochasticity) and this is why our approach is more efficient across these environments.

Understanding Failure Cases   While our approach generally exceeds the performance of or is comparable to the curiosity baselines, there are some environments where SHE underperforms. We have analyzed these games and found common failure cases: 1) Audio-visual association is trivial. For example in Qbert, the discriminator easily learns the associations: every time the Qbert agent jumps to any cube the same sound is made, thus making the discriminator's job easy, leading to a low agent reward. Visiting states with already learned audio-visual pairs is necessary for achieving a high score, even though they may not be crucial for exploration. The game Atlantis had similarly high discriminator performance and low agent rewards. 2) The game has repetitive background sounds. Games like SpaceInvaders and BeamRider have background sounds at a fixed time interval, but it is hard to visually associate these sounds. Here the discriminator has trouble learning basic cases, so the agent is unmotivated to further explore. In Alien, the agent quickly learns that by quickly passing from one side of the screen to the other, a sound occurs with a slight delay that makes it hard to align with the frame. The agent learns to repeat this trick continuously, putting the discriminator in a situation similar to 2).
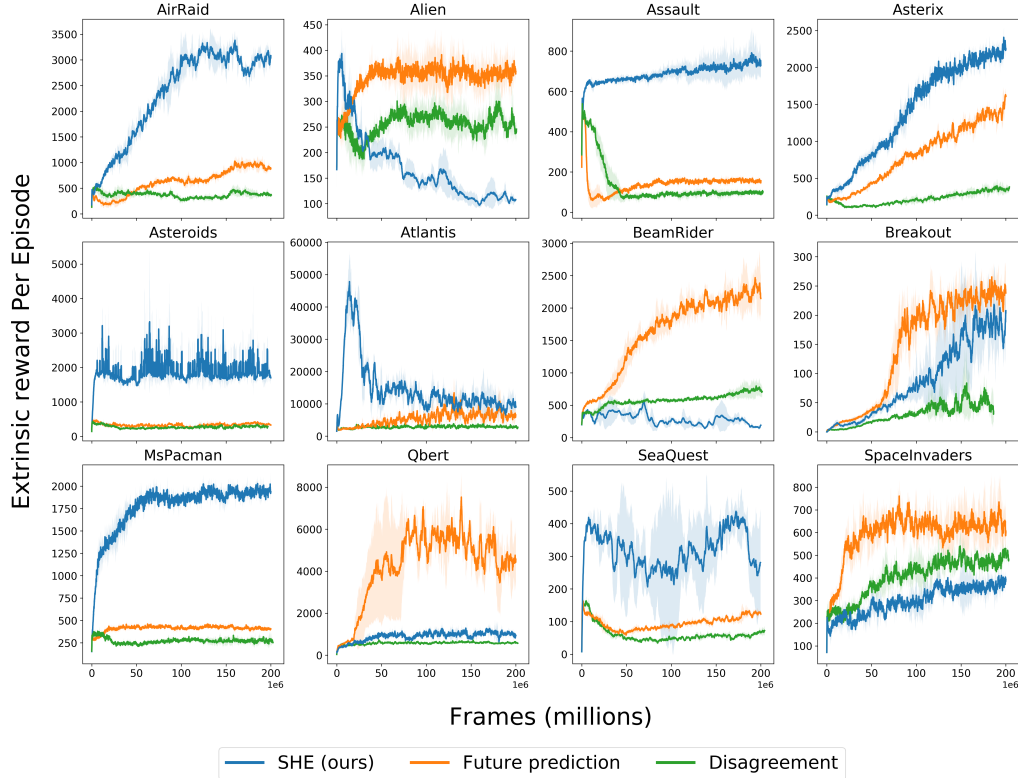
Figure 4: **Atari training curves**: Average extrinsic reward (never seen by the agent) throughout training for our method, future prediction [2], and exploration via disagreement [3]. Our method outperforms the baselines in 8 of 12 environments, supporting our hypothesis that audio-visual association is a useful signal for accelerating exploration.

Hard Exploration Environment According to Taïga et al. [41], Gravitar is a hard exploration environment. Such environments are particularly difficult to solve without learning from demonstrations [28], using extrinsic reward [41], or exploiting structure in the game [42]. Even for humans, it can be unclear how to play Gravitar upon first introduction, in contrast with other Atari games that are intuitively simple. Despite Gravitar's difficulty, SHE allows the agent to explore well, while the baselines perform poorly (Figure 5). After examining the game, we hypothesize that the game's visual dynamics are not that interesting on their own, but the audio-visual associations are. We also applied our method and the baselines to other hard exploration games, but in these cases, no method was successful in the training time allotted.
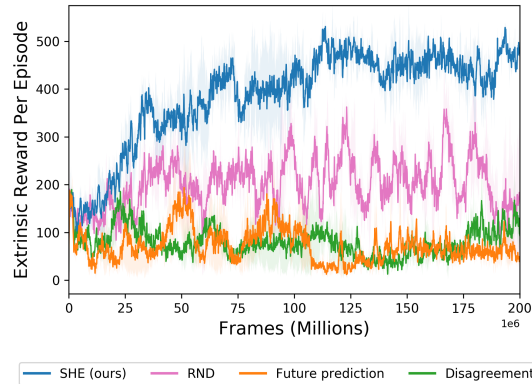

Figure 5: **Case study on Gravitar**: Our method is able to explore this hard environment, while baselines have negligible increase in extrinsic rewards.

4.4 Habitat Experimental Results

Here we present results from unsupervised area exploration in the biggest scene in Replica [40] with realistic acoustic responses [25]. Figure 6 shows the quantitative results. SHE (blue) has similar coverage to RND and reaches full state coverage 3 times faster than future prediction curiosity (Figure
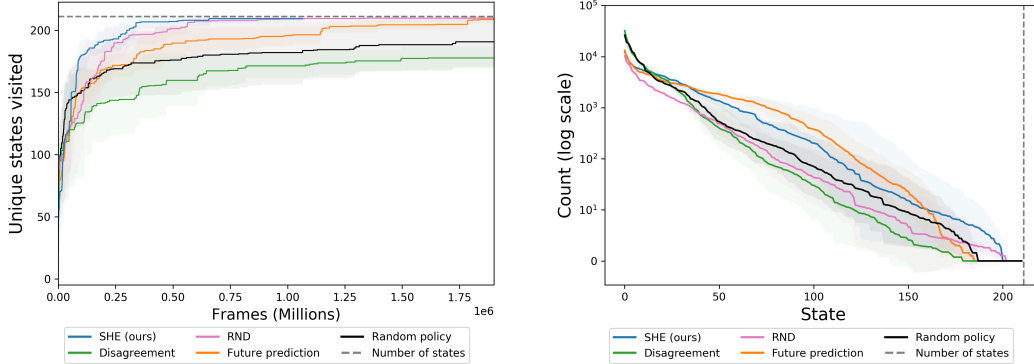
(a) State coverage, as measured by unique states visited throughout training. Our method achieves full state coverage about 3 times faster than future prediction curiosity.

(b) State counts: the number of times each state is visited in the first 2000 episodes, sorted by frequency and shown on a log scale. Our method has a wider tail, visiting rare states one to two orders of magnitude more frequently than the baselines.

Figure 6: **Habitat exploration results for SHE and baselines**. Each method is run with three different seeds and each seed uses a different start location.

6(a)). We can also look at how much each state is visited (Figure 6(b)). A good exploration method will have higher counts in the rare states. Our method visits these rare states (Figure 6(b) right) about 8 times more frequently than the next-best baseline. It does so by visiting common states (Figure 6(b) left) less frequently. SHE's strong performance on this more realistic task holds promise for future work exploring the real world.

4.5 Ablations

Audio in baseline   One hypothesis for why our method outperforms baselines is that SHE has access to additional information in the form of audio. To test the benefit of including audio without the use of our association method, we created two additional baselines: an audio-visual prediction baseline and an audio-visual random network distillation baseline. In the audio-visual prediction baseline, the prediction space is concatenated audio and visual features: the future prediction model takes an audio-visual feature vector as input and predicts an audio-visual feature vector. Similarly, in the audio-visual random network distillation baseline, the audio and visual features are concatenated and used as inputs to both the random target network and the predictor network. As the results in the appendix indicate, this does not lead to significant improvement over the visual-only baselines.

Robustness to noise   Predicting the future can be especially difficult in the face of inherent uncertainty. To analyze our approach in such a setting, we created a noisy version of the environments, where Gaussian noise is added to the audio and visual feature vector inputs. Our approach can be affected by noise in both audio and visual observations, whereas the baseline is only affected by the visual noise. For these experiments, we chose three environments: one where our method was better (MsPacman), one where the baseline was better (SpaceInvaders), and one where both methods performed well (Asterix). Figure 7 depicts results across these three environments both with and without noise. We observe that future prediction curiosity is not robust to such noise: the performance degrades significantly in both Asterix and SpaceInvaders. In contrast, as our approach only relies on associations, it is more robust to such noise.

Multiple Curiosity Modules   Curiosity can have multiple forms, e.g. prediction-based and multimodal, and these are complementary to each other. To demonstrate this, we ran a joint method combining intrinsic rewards: we sum the losses from future prediction and the audio-visual discriminator. The resulting method is better than the visual-only baseline in 10 of 12 games, sometimes surpassing both (see the appendix for the detailed results).
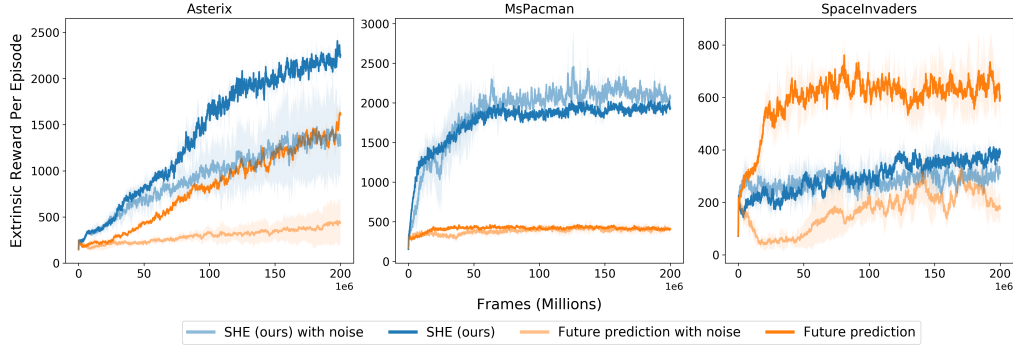
8

Figure 7: **Effect of input noise on performance**: Our method (blue) maintains similar performance with the introduction of noisy observations, while the baseline performance (orange) degrades.

5   Conclusion

Multimodality is one of the most basic facets of our rich physical world. Our formulation of curiosity enables an autonomous agent to efficiently explore a new environment by exploiting relationships between sensory modalities. With results on Atari games, we demonstrated the benefit of using audio-visual association to compute the intrinsic reward. Our method showed improved exploration over baselines in several environments. The most promise lies in our approach's significant gains when used on a more realistic task, exploration in the Habitat environment, where audio and visual are governed by the same physical processes. We anticipate multimodal agents exploring in the real world and discovering even more interesting associations. Instead of building robots that perform like adults, we should build robots that can learn the way babies do. These robots will be able to explore autonomously in real-world, unstructured environments.

Broader Impact

The lasting impact of RL will be from these algorithms working in the real world. As such, our work is centered around increasing sample efficiency and adaptability. By leveraging self-supervision, we can avoid cumbersome reward shaping, which becomes exponentially more difficult as tasks grow more complex. Although our work here uses simulated agents, our longer-term goal is to deploy multimodal curiosity on physical robots, enabling them to explore in a more sample-efficient manner. Multimodal learning could have a near-immediate impact in autonomous driving, where different sensory modalities are used for perception of near, far, small, and large entities.

Autonomous RL agents have many potential positive outcomes, such as home robots aiding elderly people or those with disabilities. They will save time and money in many sectors of industry. However, they also have the potential to displace parts of the workforce [43].

There could be privacy concerns if merged multimodal data is hard to anonymize or de-identify. There could also be privacy concerns with respect to recording audio data in the wild [44]. With unsupervised RL, it can be hard to predict what behaviors will be learned. For example, a robot using our algorithm might learn to damage sensors to create novel associations. The inability to predict agent behavior can make ensuring safety difficult, which would have consequences in safety-critical settings like autonomous driving or healthcare. Some work has been done on safety in RL [45], and there is more to be done, especially on analyzing the safety of RL exploration policies during training.

Acknowledgements

References

[1] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

[2] 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*, 2018.

[3] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, 2019.

[4] J. McV. Hunt. Intrinsic motivation and its role in psychological development. In *Nebraska Symposium on Motivation*, volume 13. University of Nebraska Press, 1965.

[5] William N Dember and Robert W Earl. Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological Review*, 64(2), 1957.

[6] Teresa Wilcox, Rebecca Woods, Catherine Chapa, and Sarah McCurry. Multisensory exploration and object individuation in infancy. *Developmental psychology*, 43(2), 2007.

[7] Ross Flom and Lorraine E Bahrick. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1), 2007.

[8] Barbara A Morrongiello, Kimberley D Fenwick, and Graham Chance. Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development*, 21(4), 1998.

[9] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: a platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[10] Katsumi Watanabe and Shinsuke Shimojo. When sound affects vision: effects of auditory grouping on visual motion perception. *Psychological Science*, 12(2), 2001.

[11] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017.

[12] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 1991.

[13] Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *CoRR*, abs/1507.00814, 2015.

[14] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks. *CoRR*, abs/1605.09674, 2016.

[15] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3), 2012.

[16] Jürgen Schmidhuber. Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, 1991.

[17] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 2016.

[18] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

[19] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

[20] Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. Loss is its own reward: Self-supervision for reinforcement learning. In *International Conference on Learning Representations*, 2017.

[21] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, 2012.

[22] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016.

[23] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[24] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[25] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[26] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[27] Ruohan Gao, Changan Chen, Ziad Al-Halab, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In *Advances in Neural Information Processing Systems*, 2018.

[29] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4), 2018.

[30] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. *International Symposium on Experimental Robotics*, 2018.

[31] Samuel Clarke, Travers Rhodes, Christopher G Atkeson, and Oliver Kroemer. Learning audio feedback for estimating amount and flow of granular material. In *Conference on Robot Learning*, 2018.

[32] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[33] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2), 2007.

[34] Adam White, Joseph Modayil, and Richard S Sutton. Surprise and curiosity for big data robotics. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[35] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2017.

[37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[39] Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018.

[40] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[41] Adrien Ali Taïga, William Fedus, Marlos C Machado, Aaron Courville, and Marc G Bellemare. Benchmarking bonus-based exploration methods on the arcade learning environment. In *International Conference on Learning Representations*, 2020.

[42] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

[43] Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370), 2017.

[44] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

[45] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), 2015.

A Code

Our code is available at: `https://github.com/vdean/audio-curiosity`

We use the following open-source repositories for baselines:
Future prediction code: `https://github.com/openai/large-scale-curiosity`
Disagreement code: `https://github.com/pathak22/exploration-by-disagreement`
RND code: `https://github.com/openai/random-network-distillation`

B Implementation Details

Input preprocessing   We convert all images to grayscale and resize to 84x84. We stack the 4 most recent frames, leading to an observation shape of 84x84x4. In Atari experiments, we use a frameskip that repeats each action 4 times. In Atari, the audio input is 530 samples per timestep, so with the frameskip and frame stack, the audio input is $530 * 4 = 2120$. In Habitat, the audio is 66150 samples per time step (each audio clip is padded to 1 second).

Network architectures   The visual embedding network is the same as that used in large scale curiosity [2]. It is a convolutional neural network with fixed random weights. It has 3 fully convolutional layers and a 512-dimensional output. The discriminator network consists of 2 densely connected layers, each with 512 hidden units, followed by another densely connected layer with a single output neuron, representing the likelihood of alignment.

Hyperparameters   In Atari, we use 128 parallel environments, and in Habitat, we use 1 environment, as it does not support multithreading. We use the same hyperparameters as in large scale curiosity: a learning rate of 0.0001 for all models, a discount factor γ of 0.99, and 3 optimization epochs per rollout.

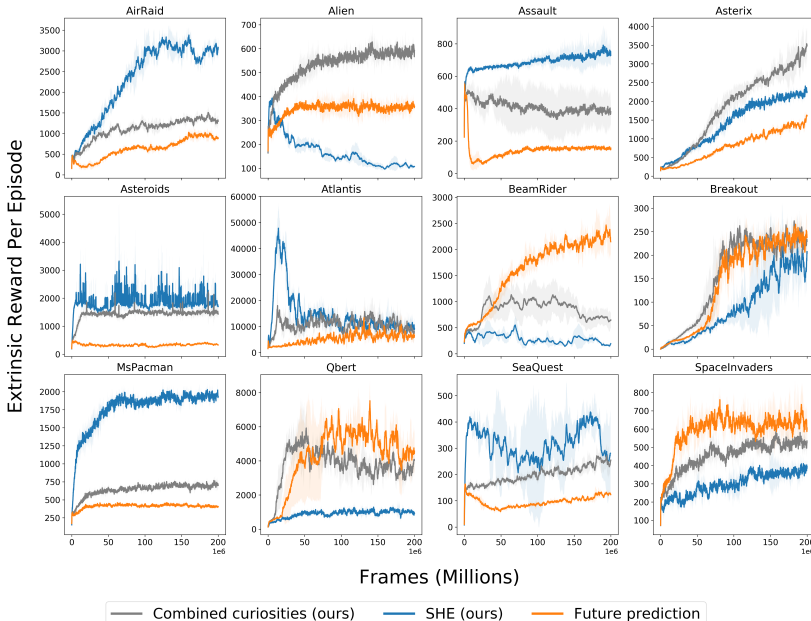## C Combining Multiple Forms of Curiosity

Figure 8: **Complementary forms of curiosity**: Maximizing the sum of SHE and future prediction intrinsic rewards leads to better exploration than future prediction alone on 10 of 12 environments.

Future prediction and multimodal association can be complementary forms of curiosity. To demonstrate this, we present a joint method in which the agent receives combined intrinsic rewards: we sum the losses from future prediction and the audio-visual discriminator. We show results using this joint method in Figure 8.

The resulting combined method is better than the future prediction baseline in 10 of 12 games, sometimes surpassing both (e.g. an average final reward of 593 compared with 108 for SHE and 359 for future prediction in Alien). Further work could explore other ways of combining intrinsic rewards, such as switching between the complementary forms. However, our goal in this work was to introduce our new approach, rather than to show strict performance improvements on every environment by tuning a combined method.

D   Using Audio in Baselines

Here we present results on using audio in baselines, as described in the main paper ablations section. In the first baseline, the prediction space is concatenated audio and visual features: the intrinsic model takes an audio-visual feature vector as input and predicts an audio-visual feature vector as output. The results from the audio-visual prediction baseline are shown in Figure 9.
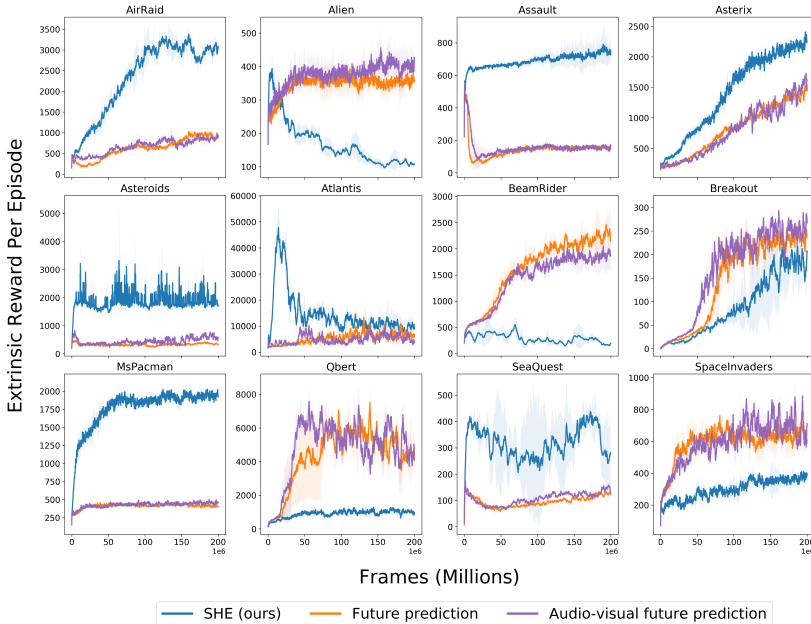
Figure 9: **Prediction baseline with audio**: The audio-visual prediction baseline performs similarly to the visual-only baseline, showing that association, rather than simple multimodality, is the key to our method's success.

In the second baseline, we add audio to random network distillation [35]. The audio and visual features are fed through separate neural network layers (dense and convolutional, respectively), concatenated, and used in both the random target network and the predictor network. The results from the audio-visual random network distillation baseline are shown on Gravitar in Figure 10.

These results demonstrate that our method's success is not just because it has additional environment information (audio). The differing sparsities and magnitudes of audio and visual features make it difficult to use both in this way. Meanwhile, our association-based method works well without additional tuning or scaling. As described in the paper, learning association instead of prediction is effective for exploration.
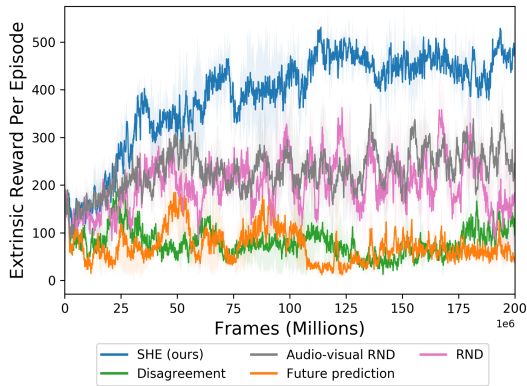
Figure 10: **Random Network Distillation baseline with audio**: On Gravitar, the audio-visual RND baseline performs similarly to the visual-only RND baseline.

14

E   Atari with Sticky Actions

While in the main Atari experiments (Section 4.3) we test on similar environments to Burda et al. [35], subsequent works such as Pathak et al. [3] have advocated for the use of a stochastic version of these environments. To test whether our method yields similar benefits when faced with the addition of stochasticity, we ran experiments using sticky versions of each environment.
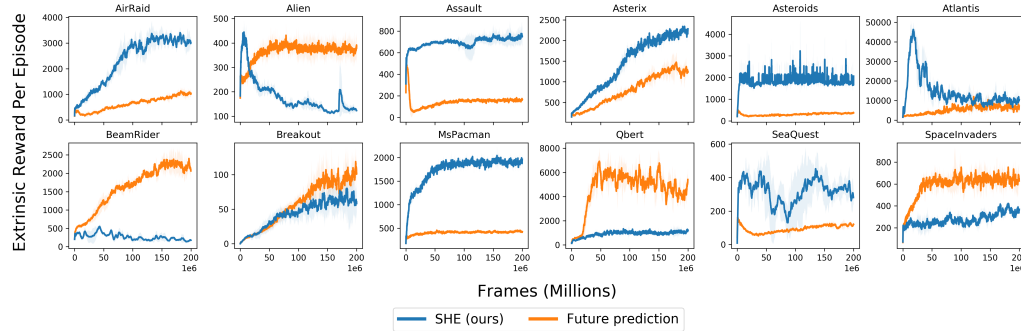

Figure 11: **Performance on stochastic environments**: In the face of action randomness, our association-based curiosity shows similar performance to the deterministic experiments.

In these "sticky" environments, the agent's action is "stuck" and repeated with probability 0.25 at each time step. These non-deterministic environments can be more challenging for the prediction-based curiosity, as the environment's added randomness makes it harder to predict. Our results, as shown in Figure 11, indicate similar trends to those observed in the non-sticky environments.

F   Habitat Exploration Visualizations

In the main paper, we presented quantitive results on the Habitat navigation setting. Qualitatively, we can look at heatmaps for where the Habitat agents explore in the scene (Figure 12). We see that baseline agents have poorer visitation in some areas far from where the agent starts (start location marked by arrows). The SHE agent has more uniform coverage, demonstrated by the more consistent shading throughout the map.
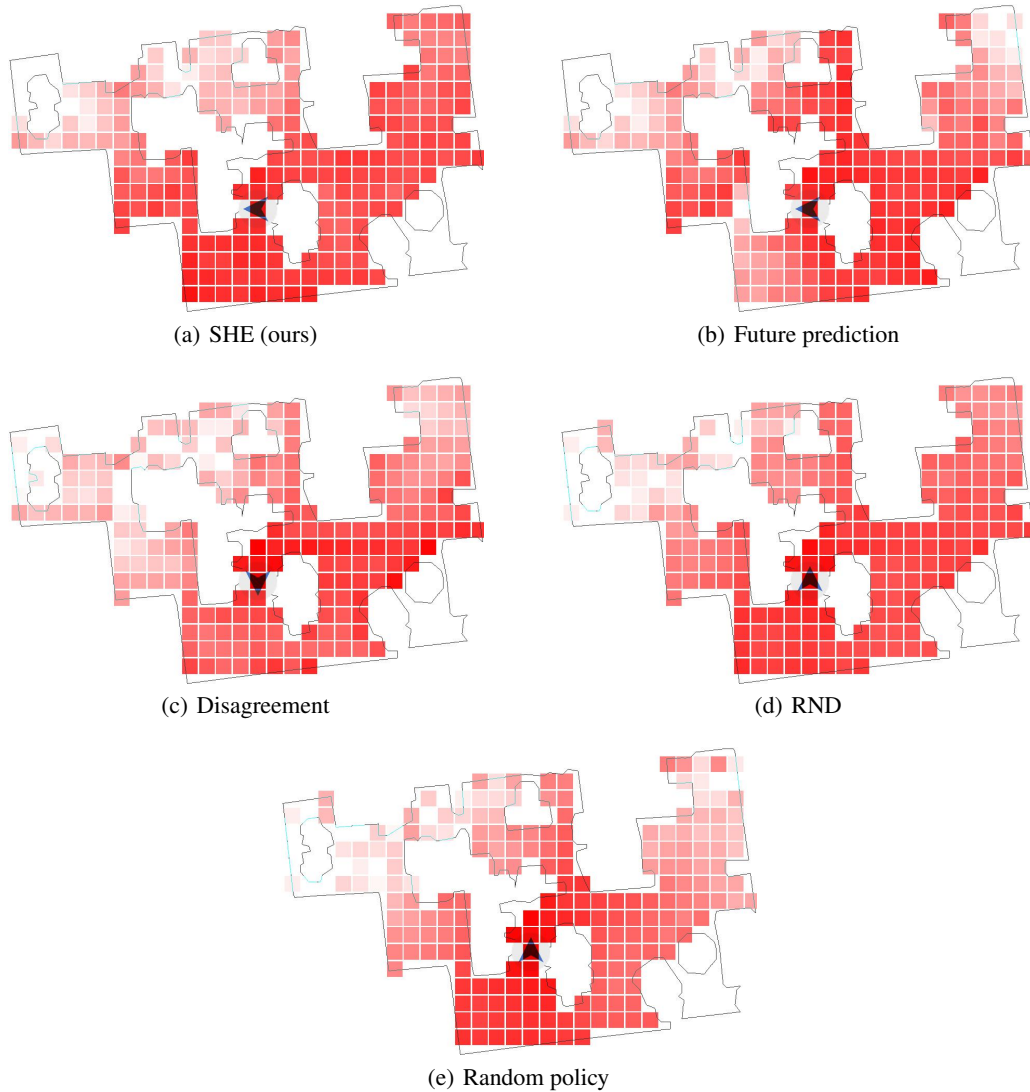
(a) SHE (ours)

(b) Future prediction

(c) Disagreement

(d) RND

(e) Random policy

Figure 12: **Navigation heatmaps for our method and the baselines**. Darker represents a higher visitation rate of the state. Visitation rates are computed using the first 2000 training episodes. Our method has more uniform visitation: it visits rarer states more and closer states less, representing better exploration. Heatmaps are made using episodes from one training seed, in which the agent started at the location marked by the arrow.