

Learning to View: Decision Transformers for Active Object Detection

Wenhao Ding^{1*,2}, Nathalie Majcherczyk¹, Mohit Deshpande¹, Xuewei Qi¹,
Ding Zhao², Rajasimman Madhivanan¹, Arnie Sen¹

Abstract—Active perception describes a broad class of techniques that couple planning and perception systems to move the robot in a way to give the robot more information about the environment. In most robotic systems, perception is typically independent of motion planning. For example, traditional object detection is passive: it operates only on the images it receives. However, we have a chance to improve the results if we allow planning to consume detection signals and move the robot to collect views that maximize the quality of the results. In this paper, we use reinforcement learning (RL) methods to control the robot in order to obtain images that maximize the detection quality. Specifically, we propose using a Decision Transformer with online fine-tuning, which first optimizes the policy with a pre-collected expert dataset and then improves the learned policy by exploring better solutions in the environment. We evaluate the performance of proposed method on an interactive dataset collected from an indoor scenario simulator. Experimental results demonstrate that our method outperforms all baselines, including expert policy and pure offline RL methods. We also provide exhaustive analyses of the reward distribution and observation space.

I. INTRODUCTION

In recent years, robot perception has improved considerably through advances in Deep Learning (DL) [1]. Static datasets, such as collections of Internet images [2], enabled the training of large and complex DL models [3]. Traditionally, these models take in images with ad-hoc views of objects and output corresponding predictions. However, this is a passive approach to robot perception and it fails to leverage the mobility of robots. The active approach to robot perception closes the loop between robot motion and perception; it involves moving the robot to acquire sensor inputs that the perception model performs well on [4]. In this paper, we specifically focus on the task of active object detection, which is illustrated in Figure 1.

To tackle the problem of active object detection, recent works [5], [6] combine object detection models with Reinforcement Learning (RL) into a decision-making framework that discovers good views by moving around a selected object [7]. In these settings, the RL agent maximizes a long-horizon objective by either learning from scratch in interactive environments or learning from pre-collected datasets. The latter, offline RL [8], is particularly suitable for robotic tasks since expert datasets can be easily obtained from demonstration or rule-based policies [9].

The major limitation of existing work comes from the Markov assumption used in traditional RL, which makes

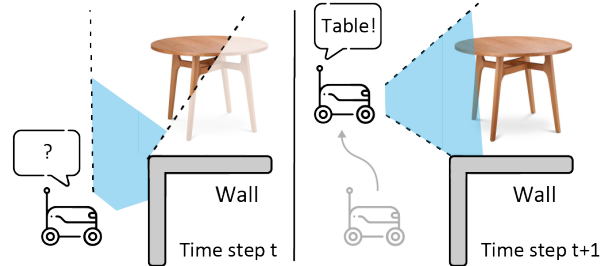


Fig. 1: An example of active object detection. At time step t , the robot cannot detect the object. After actively moving to a new position, it can accurately detect the table.

robots forget historical information that is important for memorizing visited positions.

Language models, such as recurrent neural networks and transformers, have shown great success in the natural language processing (NLP) community because of their ability to reason about *sequences* of words rather than traditional unigram or bigram models which only reason about one or two words at a time. Decision Transformers (DT) [10] are sequential generative models that apply this notion of reasoning about sequences of words to reasoning about tuples of (state, action, rewards) in an RL context. Although many problems in RL can be formulated as a Markov Decision Process (MDP), DT shows that breaking the Markov assumption to reason about sequences of tuples outperforms traditional methods.

For active object detection, previous work relies heavily on RGB images which do not provide any spatial navigation information or information about the obstacle configuration [11]. Incorporating some depth-based representation, such as a depth image, into the state would provide this missing information to the model to avoid obstacles and occlusions. Additionally, previous work suffers from training instability since they attempt to jointly optimize for the detector and the motion policy [12].

Furthermore, models trained using only offline RL methods do not generalize to out-of-distribution states and have difficulties reaching the optimal policy. In this paper, we propose to combine online and offline training on a motion policy that traverses favorable views. We first leverage expert data to pre-train the policy model in an offline manner and then fine-tune it using an interactive synthetic dataset. This type of dataset consists of a set of views collected from robot poses throughout the environment. We also propose a new sampling procedure to increase the exploration of DT for better online fine-tuning. To evaluate our proposed method, we collect an interactive

*Work done during the internship at Amazon. ¹Amazon Lab126, Sunnyvale, CA 94098, USA. ²Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: {wenhaod, dingzhao}@andrew.cmu.edu, {majcherc, deshmo, qixuewei, rajasimm, senarnie}@amazon.com

dataset from AI2THOR [13] and use it for online fine-tuning and expert dataset collection. Experiment results demonstrate that our method dramatically outperforms the expert policy.

Our contributions can be summarized as follows:

- We build upon existing online DT work [14] to tackle the active object detection problem. Our approach combines offline pre-training and online fine-tuning of a DT for active vision. We use a novel sampling method in the training of the algorithm to increase RL exploration.
- We collect an interactive dataset in the AI2THOR simulator based on the protocol and detection model from [12], which contains numerous scenes for the active object detection task.
- We provide exhaustive analyses of our proposed method and a thorough comparison with strong baselines.

We organize this paper as follows. We first discuss related work in Section II. Then, we formalize the problem statement in Section III. The design of our method is presented in Section IV. We report the results of our performance evaluation in Section V, and conclude the paper in Section VI.

II. RELATED WORK

A. Active Object Detection

Embodied AI [15] is an emerging topic that learns through interactions with the environment from an egocentric perception instead of learning from static datasets of images, videos, or text from the internet. Active object detection is one typical task of embodied AI and has been studied for many years. Ammirato et al. [5] propose a benchmark for active vision tasks that includes an interactive dataset. They implement a baseline approach that uses the REINFORCE [16] algorithm with the object detection score as the reward. A following work [17] proposes to use DQN [18] to solve this problem. [6] also relies on RL but adds an aggregator module to accumulate historical information to find the next best view. [11] uses an occlusion ratio as the reward to train the motion policy and proposes a three-stage pipeline. They first train the detection model, then train the motion policy, and finally fine-tune the detection model to achieve better performance. [19] collects multi-view data, generates pseudo-labels, and fine-tunes the object detector in a self-supervised manner.

Besides the RL formulation, a meta-learning framework [20] is also used to solve the active object detection task [12], [21]. Instead of maximizing a reward that indicates the performance, they train the detection model on a sequence of images to maximize the detection results of the first frame. Thus the policy model aims to find images that can help the detection model quickly adapt to the first frame.

B. Decision Transformer

Transformers [3] achieve great success in NLP tasks, such as language translation and sentence prediction. They use the self-attention mechanism, which computes the correlation between the inputs and outputs.

Recently, researchers adapt Transformer models to decision-making tasks [10], [22] since it can be used for action prediction conditioned on historical states and reward-to-go.

Although trained in a supervised learning manner, DT outputs different actions to match different target reward-to-go. More variants of DT are proposed later, for example, [23] extends DT to stochastic environments, [24] incorporates counterfactual reasoning into DT, and [25] generalizes DT to broader usage for hindsight information matching.

C. Offline Reinforcement Learning

RL algorithms [26], [27], [28] learn from the interaction between agent and environment, usually leading to inefficient training and dangerous exploration. Although off-policy [26] algorithms make it possible to use a replay buffer for model training, they build the buffer with data collected from their online learning policy. Recently, some “truly” off-policy algorithms, also known as offline RL, build the buffer with data collected from human experts or rule-based methods [29].

Offline RL can be simply implemented [30] by adding regularization to DQNs [18], which avoids the over-estimation of unseen action-state value in the pre-collected dataset. To avoid evaluating unseen actions outside of the dataset, [31] proposes IQL that implicitly approximates the policy improvement step. More advanced variants of offline DQN are introduced in [32].

Most offline RL methods are limited by the quality of the dataset, e.g., a dataset collected from a sub-optimal policy makes the agent never see optimal states. Two solutions are proposed to break the limitation of the low-quality dataset. One is encouraging exploration during dataset generation [33] and the other is adding an online fine-tune stage after offline training [34], [14].

III. PROBLEM STATEMENT

A. Objective

The aim of active object detection is to learn a motion policy π_θ to control a mobile robot such that the predictions of the object detection model \mathcal{M}_ϕ are maximal. The detection model predicts bounding boxes \hat{b} of objects over RGB images such that $\mathcal{M}_\phi : I^{RGB} \rightarrow (\hat{b}, \delta)$, where δ is the detection confidence score associated with the prediction. δ is set to 0 if the prediction is wrong. We denote the corresponding depth image as I^D and T refers to the maximum number of steps that can be taken to improve the detection performance.

B. Assumptions

To focus on the core problem, we make several important assumptions in our framework:

- To avoid training instability, we assume the weights of the detection model \mathcal{M}_ϕ are frozen.
- We discretize the state and action spaces to simplify dataset collection.
- We use the detection score as a surrogate value for detection performance. Although previous work [35] claims that this score may not be a good metric, we find it still useful as shown in our experimental analysis.
- We assume that the robot has access to depth images, which allow us to consider environmental context and distances to surrounding objects.

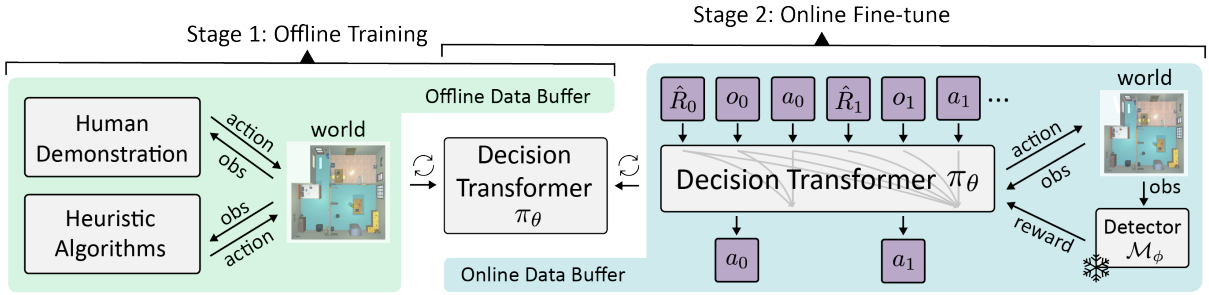


Fig. 2: Illustration of the proposed framework. In Stage 1, we conduct offline training on the Decision Transformer with expert data collected from human or heuristic algorithms. In Stage 2, we fine-tune the model in an interactive environment, where the reward comes from a detection model with frozen parameters.

C. Formulation

We formulate this problem as a Partially Observable Markov Decision Process (POMDP), where the agent cannot directly observe the underlying state. $s \in \mathcal{S}$ represents the internal state, including parameters ϕ of \mathcal{M}_ϕ and properties of the object to be detected. $o \in \mathcal{O}(o|s', a)$ represents the observable state to the policy model, which contains the bounding box of selected object \hat{b} , the RGB image I^{RGB} , and the depth image I^D . $a \in \mathcal{A}$ represents the output action of the policy, which is selected from the set $\mathcal{A} = \{\uparrow, \downarrow, \leftarrow, \rightarrow, \cup, \cap, \times\}$. Finally, the reward $r \in \mathcal{R}$ represents the performance of the detection model with $r = \delta(I^{RGB})$. Our goal is to maximize the expected reward $\max_\theta \mathbb{E}_{\pi_\theta} [\sum_{t=0}^T r_t]$, where the policy model π_θ takes as input an observation o and outputs an action a .

IV. METHODOLOGY

In this paper, we propose to tackle the Active Object Detection problem using a DT as the policy model π_θ and introduce it in Section IV-A. We also discuss the offline-online learning framework designed for efficiently training this model in Section IV-B. The entire procedure of the algorithm is summarized in Algorithm 1.

A. Policy Model Design

The key idea behind Decision Transformers is conditional action generation, where the output action is conditioned on the desired future rewards represented by the reward-to-go (RTG):

$$\hat{R}_t = \sum_{t'=t}^T r_{t'}, \quad (1)$$

where t and T are the current time step and the final time step. Decision Transformers take in as input a sequence of RTGs, observations, and actions. We refer to this sequence as trajectory τ :

$$\tau = (\hat{R}_0, o_0, a_0, \dots, \hat{R}_t, o_t). \quad (2)$$

The ground truth of the output sequence is a_t . Since we use discrete actions, we select cross entropy \mathcal{L}_{CE} as the training loss function. At inference time, the true RTG is not known so we approximate it using an estimate of the maximum accumulated reward for the task. Since the maximum reward

Algorithm 1 Offline-Online Training of DT

- 1: **Input:** Expert Policy π_e , Buffer Size N , Simulator \mathcal{E}
- 2: **Output:** Motion Policy π_θ
- 3: **Stage 1: Offline Training**
- 4: **for** i in $[1, N]$ **do**
- 5: Roll out π_e in \mathcal{E} to get trajectory τ_i , $\mathcal{B} \leftarrow \mathcal{B} \cup \{\tau_i\}$
- 6: Train π_θ with \mathcal{L} with buffer \mathcal{B}
- 7: **Stage 2: Online Fine-tuning**
- 8: **while** not converged **do**
- 9: Roll out π_θ in \mathcal{E} to get trajectory τ_i
- 10: Remove the first point in \mathcal{B} , $\mathcal{B} \leftarrow \mathcal{B} \cup \{\tau_i\}$
- 11: Train π_θ with \mathcal{L}_{CE} with buffer \mathcal{B}

for each time step is in $[0, 1]$ as we use the detection score, we set the initial RTG \hat{R}_0 to the length of episode T . Then, the RTG is updated by subtracting the true reward obtained at each step $\hat{R}_{t+1} = \hat{R}_t - r_t$.

The training of DT follows the same protocol as supervised training and thus requires batch sampling from the data buffer. In the original DT paper [10], the probability of sampling one trajectory is proportional to the length of the trajectory. We sample the data buffer proportional to the variance of the rewards in the trajectory to encourage exploration and allow the model to see a diverse set of samples. Since the variance of the states is difficult to calculate, we use the variance of the rewards as a tractable proxy. Specifically, for the i -th trajectory, the probability is

$$p_i = \frac{\text{Var}(\tau_i)}{\sum_i^N \text{Var}(\tau_i)} \quad (3)$$

where N is the total number of training trajectories. To further increase exploration, we add an entropy term \mathcal{L}_{en} of the predicted action, resulting in the final loss function:

$$\mathcal{L} = \mathcal{L}_{CE} - \lambda \mathcal{L}_{en}, \quad (4)$$

where λ is empirically set to 0.1 for balancing the weight of two training losses.

B. Offline and Online Training Framework

We illustrate the proposed offline-online training framework in Figure 2. The original DT requires a data buffer \mathcal{B} that stores the pre-collected dataset. To generate such a dataset, we design

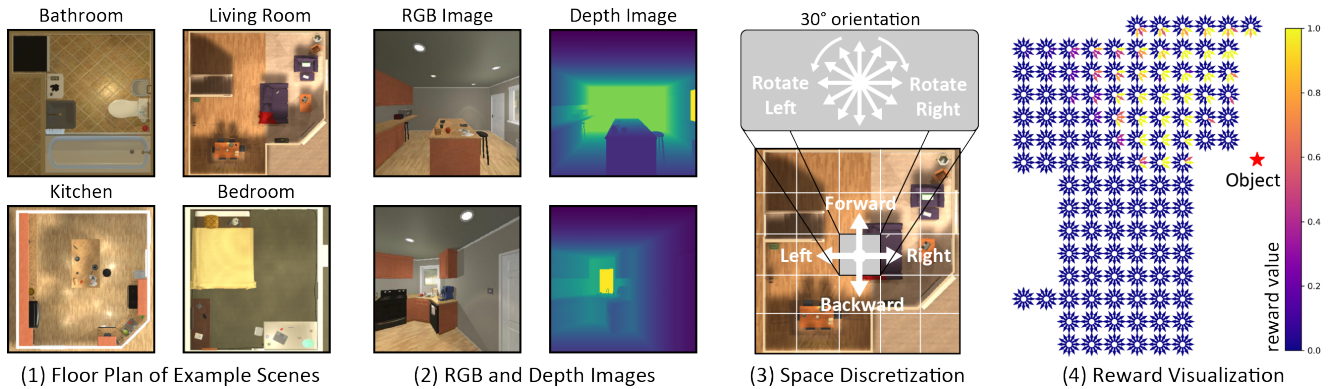


Fig. 3: Description of the interactive dataset. (1) The floor plan of 4 categories of scenes provided by the AI2THOR simulator. (2) The RGB and depth images obtained from the simulator. (3) Discretization of the location and orientation together with 7 discrete actions. (4) Reward (detection score) of one object in the discretized dataset.

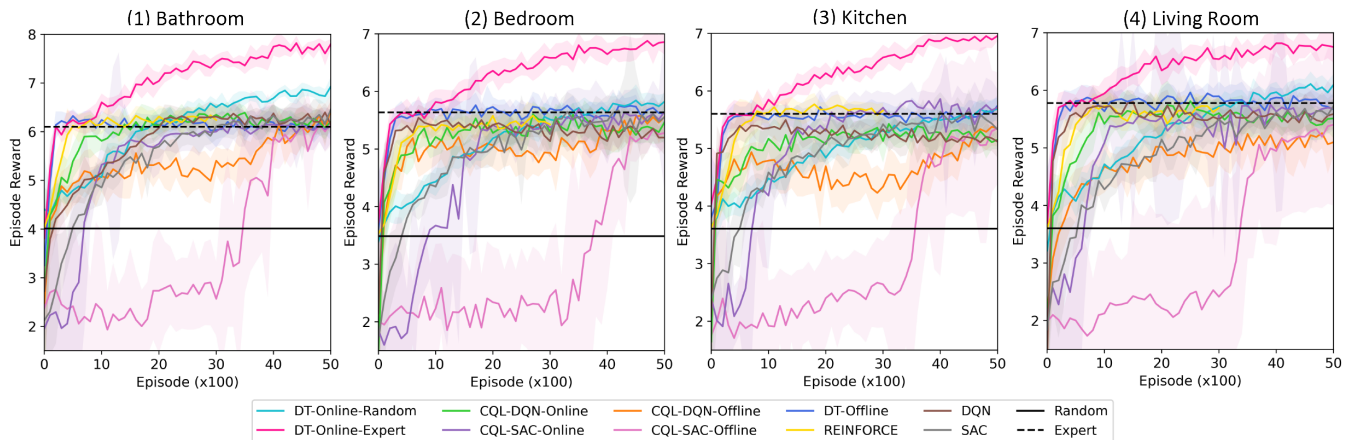


Fig. 4: Episode reward of proposed method and baselines in 4 categories of floor plans.

a heuristic expert that imitates human logic: it first rotates to put the object in the middle of the view and then goes forward to minimize the distance to the object. Note that this simple heuristic policy is by no means optimal since the selected object could be occluded by other objects in the scene. Purely training the DT on this dataset makes it almost impossible to achieve optimal performance, thus we still need to obtain new information from the interactive environment.

Therefore, we propose to add an online training stage after the offline training, which gradually replaces samples in the data buffer \mathcal{B} with new data obtained by the learned policy π_θ . After 10 episodes of online data collection, we use the new data buffer to train the policy π_θ and repeat this process until reaching the maximum number of episodes. When collecting the new data sample, we use the true RTG calculated from the reward rather than the approximated RTG. This procedure is also called Hindsight Return Relabeling, which is explored in a similar offline-online framework in [14].

V. EXPERIMENTS

We first describe the design and collection process of our interactive dataset in Section V-A. Then, we introduce the overall performance of our method compared against several baselines in Section V-C. Finally, we study the reward

distribution in Section V-D and perform an ablation study in Section V-C.

A. Interactive Dataset Collection

AI2THOR [13] is a popular simulator for embodied AI and is used by Kotar et al. [12] to collect their interactive dataset. Since we propose a new formulation of the active object detection problem, we create a new dataset. There are 4 categories of scenes (*Bathroom*, *Kitchen*, *Living Room*, and *Bedroom*) as shown in Figure 3 (1). Each scene contains 30-floor plans. The robot navigates in the 2D scene and acquires observations shown in Figure 3 (2), including RGB images and depth images.

The reason we collect this dataset instead of directly using the simulator is that we can avoid online rendering by querying stored images to speed up the interaction process. However, we cannot store all images for all continuous locations. Therefore, we discretize the space into 30 cm-sized cells as shown in Figure 3 (3). For each cell, we quantize the orientation into 30 degree increments. The action space is a set of actions that move the robot linearly or angularly by these increments. When the robot takes an action, we query the corresponding image from the dataset.

The last missing part of the dataset is the reward. As

TABLE I: Detection results of different motion policies.

Method	Bathroom			Bedroom			Kitchen			Living Room		
	Recall	Precision	Reward	Recall	Precision	Reward	Recall	Precision	Reward	Recall	Precision	Reward
Random	0.332	0.374	4.03±0.10	0.254	0.300	3.58±0.09	0.246	0.297	3.59±0.11	0.260	0.325	3.60±0.15
Expert	0.554	0.828	6.10±0.16	0.466	0.867	5.62±0.19	0.385	0.755	5.66±0.01	0.500	0.853	5.59±0.21
REINFORCE	0.567	0.832	6.24±0.21	0.441	0.837	5.43±0.18	0.376	0.764	5.66±0.14	0.532	0.869	5.70±0.17
DQN	0.570	0.826	6.15±0.23	0.402	0.815	5.20±0.30	0.381	0.742	5.12±0.28	0.461	0.827	5.59±0.26
SAC	0.535	0.814	6.45±0.30	0.442	0.861	5.55±0.47	0.373	0.731	5.33±0.23	0.487	0.823	5.73±0.23
CQL-DQN-Offline	0.535	0.824	6.44±0.19	0.456	0.852	5.56±0.29	0.331	0.693	5.35±0.31	0.320	0.754	5.10±0.23
CQL-SAC-Offline	0.563	0.837	6.17±0.46	0.418	0.824	5.28±0.24	0.314	0.703	5.35±0.97	0.481	0.832	5.41±1.40
DT-Offline	0.546	0.821	6.12±0.29	0.472	0.859	5.62±0.15	0.392	0.742	5.74±0.18	0.515	0.848	5.67±0.23
CQL-DQN-Online	0.560	0.831	6.21±0.19	0.435	0.853	5.46±0.32	0.392	0.748	5.15±0.16	0.463	0.832	5.51±0.34
CQL-SAC-Online	0.552	0.811	6.11±0.34	0.452	0.871	5.65±1.35	0.371	0.750	5.72±0.31	0.503	0.849	5.68±0.37
DT-Online-Random	0.672	0.875	6.93±0.19	0.478	0.853	5.82±0.14	0.369	0.743	5.69±0.20	0.510	0.851	6.10±0.19
DT-Online-Expert	0.734	0.901	7.80±0.10	0.785	0.907	6.86±0.21	0.692	0.876	6.96±0.08	0.779	0.898	6.75±0.27

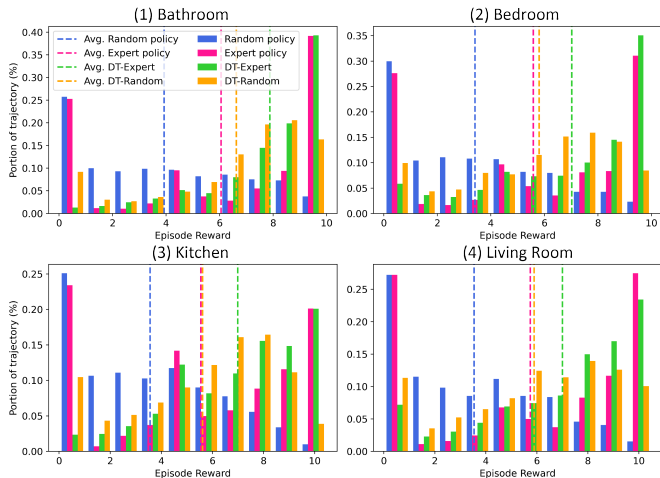


Fig. 5: Distribution of episode reward on 4 scenes. The expert policy is not optimal since it fails in $\sim 25\%$ of cases. In contrast, our method reduces the portion of failure cases to $\sim 5\%$.

introduced in Section III-A, we use the detection score of the model \mathcal{M}_ϕ as the reward. To generate such rewards, we use the DETR [36] model with pre-trained weights provided by [12]. We plot an example of the reward over the reachable space for one object in Figure 3 (4). We observe that the value of the reward is consistent with the distance between the robot and the object, and the reward drops to 0 if the robot is too close.

Not all poses are meaningful to start from. Poses where the object is already clearly in view or poses where we do not get a detection do not start useful trajectories for training the model. We apply several rules to select the initial poses: the reward is lower than a threshold of 0.2, the area of bounding box \hat{b} is larger than 200, and the distance between the object and the robot is larger than 3 m. After applying these rules, we obtain ~ 1000 initial poses in each scene category.

B. Baselines

To focus on the offline and online training framework, we consider 4 types of baseline models:

- Heuristic Algorithms. The simplest rule is randomly selecting actions from the action set, which is named **Random**. We also design a human-inspired algorithm

named **Expert** that first rotates to put the object in the middle of the view and then moves forward.

- Online RL. **DQN** [18] and **SAC** [26] are two well-known off-policy RL methods that can be used with a discrete action space. **REINFORCE** [16] is used by Ammirato et al. [5] as the motion policy.
- Offline RL. CQL [30] is a state-of-the-art offline RL method, which is implemented based on either DQN (**CQL-DQN-Offline**) or SAC (**CQL-SAC-Offline**). We also consider the original version **DT-Offline** as a baseline. All these methods are trained offline with a data buffer collected by the expert policy.
- Offline RL with Online Training. We adapt the DQN and SAC flavors of CQL for online training and name them **CQL-DQN-Online** and **CQL-SAC-Online**. **DT-Online-Random** is a special case that is first trained with a data buffer collected by the random policy and then fine-tuned online.

All these methods are implemented in PyTorch [37]. The training of motion policy runs on an NVIDIA RTX 3080Ti GPU with 12 GB memory.

C. Performance Evaluation

We show the episode reward results of our method (**DT-Online-Expert**) and all baselines in Figure 4. We note that our method outperforms all other methods including the **Expert** method. Compared with DT-Online-Random, our method bootstraps from the expert policy and outperforms it through online interaction with the environment. DT-Offline only bootstraps from the expert policy without collecting additional samples from the environment. Therefore, the advantage of DT-Online-Expert comes from both offline pre-training and online fine-tuning. Lastly, although CQL-DQN-Online and CQL-SAC-Online share the same training framework as ours, CQL fails to achieve the same performance as our method. Intuitively, this indicates that the Markov assumption fails to capture the complexity of this problem and that temporal information benefits the training of the motion policy.

Since we use the detection score as a surrogate metric for the performance of the object detection model, we need to validate that this reward captures the real objective: maximizing the detection quality. We compute the precision and recall of the

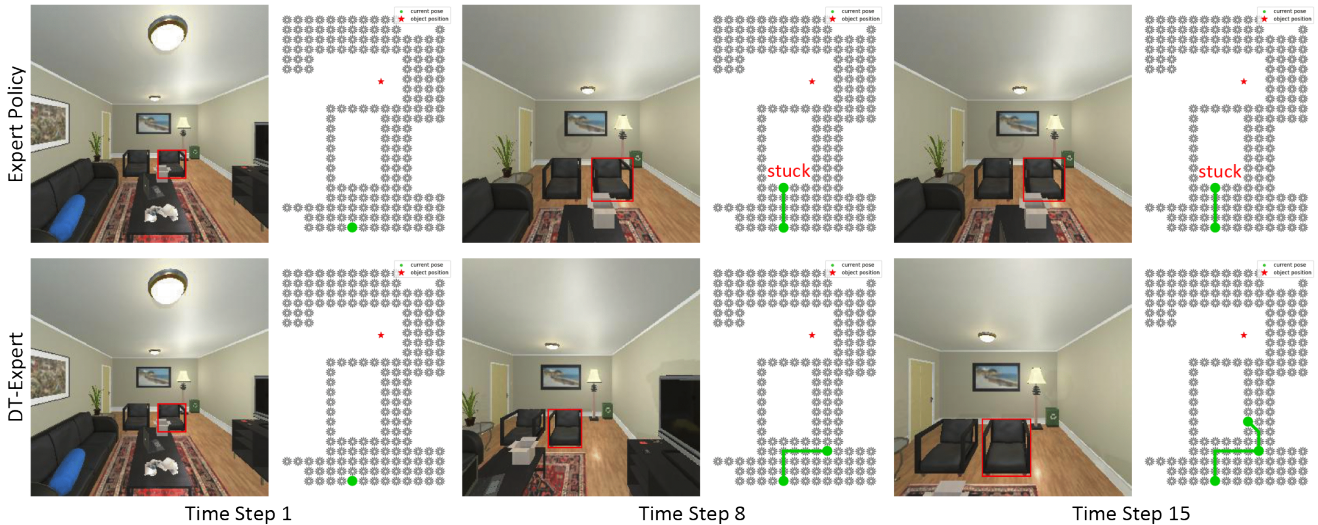


Fig. 6: Challenging example with an armchair as the selected object. The top row shows the trajectory generated by the expert policy and the bottom row shows the trajectory generated by DT-Expert. The expert is stuck by the table while the DT-Expert bypasses the table and finds a good view of the armchair.

detector along the trajectory (Table I) and show that our method still achieves the best results, which indicates that using the detection score is reasonable and the motion policy indeed improves the detection quality.

D. Reward Distribution

To further analyze the performance of our approach, we look into which scenarios can be solved by our method but not by the expert. We plot the distribution of episode rewards from 4 different methods in Figure 5. The x-axis is the episode reward and the y-axis is the portion of scenarios.

Unsurprisingly, the distribution of rewards of the random policy concentrates on the low-reward region and receives a reward of 0 in $\sim 25\%$ of scenarios. However, we notice that the expert policy receives a reward of 10 in $\sim 25\%$ of scenarios but also a reward of 0 in another $\sim 25\%$ of scenarios, which indicates that the expert policy has a lot of failure cases.

Now let us turn to learning-based methods. DT-Random reduces the number of low-reward scenarios but cannot achieve high rewards in most scenarios. Our DT-Expert method further reduces the number of low-reward scenarios and perfectly solves most scenarios. In summary, DT-Expert outperforms the baselines because it solves some challenging cases that cannot be solved by other methods. We show one example of the challenging cases in Figure 6 for better illustration. We find that Expert fails because the robot is stuck by an obstacle in the scene. In contrast, DT-Expert bypasses this obstacle using the depth information from depth images.

TABLE II: Ablation study of observation space.

Scenario	w/o Both	w/o Depth	w/o RGB	Depth+RGB
Bathroom	6.10 \pm 0.14	6.14 \pm 0.17	7.75 \pm 0.12	7.80 \pm 0.10
Bedroom	5.72 \pm 0.17	5.70 \pm 0.21	6.84 \pm 0.13	6.86 \pm 0.21
Kitchen	5.58 \pm 0.08	5.61 \pm 0.05	6.95 \pm 0.15	6.96 \pm 0.08
Living Room	5.52 \pm 0.13	5.52 \pm 0.18	6.75 \pm 0.21	6.75 \pm 0.27

E. Ablation study

To investigate the influence of the observation space on the DT model, we conduct two ablation experiments, one which removes the RGB image and one which removes both the RGB image and the depth image. We notice that the RGB image provides limited information compared to the depth image since active object detection heavily relies on the distance between the robot and the object as well as the obstacles in front of the robot. Since the depth image provides more information than the RGB image, if we had a local map around the robot, we suspect that would provide even more spatial information to the DT model to better solve the task.

VI. CONCLUSIONS

In this paper, we focused on the active object detection problem, where a motion policy is trained to control the robot to achieve better detection performance. To efficiently investigate this problem, we collected an interactive dataset in the AI2THOR simulator, used the detection score of a pre-trained detection model as the reward, and designed several rules for initial pose selection. Then, we applied Decision Transformers to this problem and proposed a novel offline-online training framework. We demonstrated the proposed method efficiently combines the advantages of offline expert information and online interactions with the environment. Further analysis of scenario difficulty and ablation studies illustrated that the spatial context is important to avoid obstacles and find the best view of the selected object.

There are still critical limitations of the proposed method. We use the detection score to indicate the detection performance, and this metric is unreliable for out-of-distribution inputs to the detection model. Finding other metrics to measure the uncertainty of detection will be an important future direction. Stably training the detection and motion models is another avenue to explore.

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [4] R. Zeng, Y. Wen, W. Zhao, and Y.-J. Liu, "View planning in robot active vision: A survey of systems, algorithms, and applications," *Computational Visual Media*, vol. 6, no. 3, pp. 225–245, 2020.
- [5] P. Ammirato, P. Poirson, E. Park, J. Kořecká, and A. C. Berg, "A dataset for developing and benchmarking active vision," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1378–1385.
- [6] D. Jayaraman and K. Grauman, "End-to-end policy learning for active visual categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1601–1614, 2018.
- [7] W. Ding, B. Chen, M. Xu, and D. Zhao, "Learning to collide: An adaptive safety-critical scenarios generating method," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2243–2250.
- [8] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.
- [9] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, "D4rl: Datasets for deep data-driven reinforcement learning," *arXiv preprint arXiv:2004.07219*, 2020.
- [10] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.
- [11] J. Yang, Z. Ren, M. Xu, X. Chen, D. J. Crandall, D. Parikh, and D. Batra, "Embodied amodal recognition: Learning to move to perceive objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2040–2050.
- [12] K. Kotar and R. Mottaghi, "Interactron: Embodied adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 860–14 869.
- [13] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.
- [14] Q. Zheng, A. Zhang, and A. Grover, "Online decision transformer," *arXiv preprint arXiv:2202.05607*, 2022.
- [15] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.
- [16] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.
- [17] X. Han, H. Liu, F. Sun, and X. Zhang, "Active object detection with multistep action prediction using deep q-network," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3723–3731, 2019.
- [18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
- [19] Z. Fang, A. Jain, G. Sarch, A. W. Harley, and K. Fragkiadaki, "Move to see better: Self-improving embodied object detection," *arXiv preprint arXiv:2012.00057*, 2020.
- [20] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [21] M. Wortsman, K. Ehsani, M. Rastegari, A. Farhadi, and R. Mottaghi, "Learning to learn how to learn: Self-adaptive visual navigation using meta-learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6750–6759.
- [22] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.
- [23] K. Paster, S. McIlraith, and J. Ba, "You can't count on luck: Why decision transformers fail in stochastic environments," *arXiv preprint arXiv:2205.15967*, 2022.
- [24] A. Boustati, H. Chockler, and D. C. McNamee, "Transfer learning with causal counterfactual reasoning in decision transformers," *arXiv preprint arXiv:2110.14355*, 2021.
- [25] H. Furuta, Y. Matsuo, and S. S. Gu, "Generalized decision transformer for offline hindsight information matching," *arXiv preprint arXiv:2111.10364*, 2021.
- [26] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.
- [27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [28] W. Ding, H. Lin, B. Li, and D. Zhao, "Generalizing goal-conditioned reinforcement learning with variational causal reasoning," *arXiv preprint arXiv:2207.09081*, 2022.
- [29] R. F. Prudencio, M. R. Maximo, and E. L. Colombari, "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *arXiv preprint arXiv:2203.01387*, 2022.
- [30] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1179–1191, 2020.
- [31] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," *arXiv preprint arXiv:2110.06169*, 2021.
- [32] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 104–114.
- [33] D. Yarats, D. Brandfonbrener, H. Liu, M. Laskin, P. Abbeel, A. Lazaric, and L. Pinto, "Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning," *arXiv preprint arXiv:2201.13425*, 2022.
- [34] A. Nair, A. Gupta, M. Dalal, and S. Levine, "Awac: Accelerating online reinforcement learning with offline datasets," *arXiv preprint arXiv:2006.09359*, 2020.
- [35] D. Gallos and F. Ferrie, "Active vision in the era of convolutional neural networks," in *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE, 2019, pp. 81–88.
- [36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.