Hindsight Relabeling is All You Need for Reach-Avoid Learning

Kevin Li
Massachusetts Institute of Technology
Harvard Medical School
likevin@mit.edu

Marinka Zitnik
Harvard Medical School
Kempner Institute at Harvard University
marinka@hms.harvard.edu

Abstract

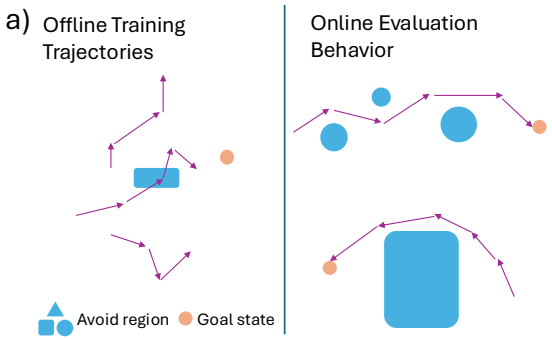
Offline goal-conditioned reinforcement learning methods have shown promise for reach-avoid tasks, where an agent must reach a target state while avoiding undesirable regions of the state space. Existing approaches typically encode avoid-region information into an augmented state space and cost function, which prevents flexible, dynamic specification of novel avoid-region information at evaluation time. They also rely heavily on meticulously designed reward and cost functions, limiting transferability to novel environments and specifications. We introduce RADT, a decision transformer model for offline, reward-free, goal-conditioned, avoid-region-conditioned RL. RADT encodes goals and avoid-regions directly as prompt tokens, allowing any number of avoid-regions of arbitrary size to be specified at evaluation time. Using only suboptimal offline trajectories from a random policy, RADT can learn reach-avoid behavior in a completely data-driven manner using our novel avoid-region hindsight relabeling approach, without the need for reward/cost supervision. We benchmark RADT against offline goal-conditioned RL models on 17 tasks that span multiple environments and experimental settings. RADT generalizes zero-shot to avoid-region sizes and counts outside the training distribution and outperforms baselines that require retraining. In one zero-shot setting, RADT reduces normalized cost by 35.7% relative to the best retrained baseline and achieves a high goal-reaching success rate. We also apply RADT to cell reprogramming, where the model plans perturbation sequences that move cells toward target cell types and away from undesirable intermediate cell states.

1 Introduction

Many sequential decision-making problems [35, 17, 1] are naturally framed as reach-avoid tasks [21, 50, 11], in which an agent must reach a designated goal state while avoiding undesirable regions of the state space. These problems arise in diverse domains, including robotics [17, 47, 6] (e.g., robotic arms reaching for targets while avoiding fragile objects), vehicle navigation [35, 1] (e.g., routing to a destination while avoiding slow zones), and biology [61, 34] (e.g., cell reprogramming strategies that reach a desired gene expression state without traversing tumorigenic intermediates). Despite domain-specific differences, these tasks share a common structure: they require balancing goal achievement with dynamic avoidance of specified hazards.

A diverse array of tasks fall under the reach-avoid framework, each with varying degrees of tolerance to goal-reaching and avoidance success that impose assorted sets of challenges, restrictions, and requirements. In safety-critical tasks where entering undesirable states can have irreversible and unacceptable consequences (e.g., an autonomous vehicle that must avoid pedestrians), online exploration is often precluded, making offline learning necessary [35]. In tasks where undesirable regions are not hard restrictions and could even be unavoidable (e.g., a vehicle must reach its destination, but should try to avoid construction zones for greater efficiency), methods derived under the framework of stringent constraint satisfaction may not be suitable [2, 3, 62]. Furthermore, in practical deployments, the specification of goals and avoid-regions may change based on user preferences or

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



b)

	Offline	Goal-conditioned	Avoid region-conditioned	Reward-free/ cost-free
OGCRL	√	√	x	√
OSRL	√	x	x	x
OGCRL + OSRL (RbSL, AM-Lag)	√	^	^	x
RADT	√	√	√	√

√ Fully satisfied ^ Partially satisfied x Not satisfied

Figure 1: (a) An ideal reach-avoid model should learn to avoid arbitrarily specified regions of varying number and size at evaluation time, using only suboptimal, random-policy training data. (b) RADT is the only model that satisfies all criteria for an ideal reach-avoid learner (Section 2).

environmental context. For instance, a robot assistant may need to adapt to new furniture layouts, or a therapeutic model may need to avoid different toxic intermediate states based on patient-specific risk factors. These settings require flexible and interpretable models that support zero-shot generalization to unseen goal and avoid specifications without retraining.

Due to these challenges, reach-avoid learning remains difficult. Several lines of work attempt to address parts of the reach-avoid problem (Section 3), but fail to meet one or more key criteria needed for flexible and effective reach-avoid learning, including: offline learning from suboptimal data, dynamic test-time conditioning on arbitrary goals/avoid-regions, and reward-free training (Section 2, Figure 1). Most existing approaches rely on augmented state representations and carefully shaped cost functions to encode avoid behavior [6, 62, 68, 30, 29, 35]. This tight coupling of avoid-region semantics to model internals prevents flexible deployment and limits generalization. Reward-based formulations often struggle to represent multiple behavioral preferences simultaneously, especially when goal-reaching and avoidance conflict, and rely on constrained optimization setups that imply an inherent prioritization of avoidance over goal-reaching that is not always desired [1, 12, 25, 2, 3, 62]. Reward-free methods avoid these issues but lack a mechanism for dynamically conditioning behavior on new avoid constraints [15, 64, 22, 10]. Moreover, many offline approaches rely on expert near-optimal demonstrations to learn strong policies [43, 35, 6, 13, 27], which are often not available in safety-critical or high-dimensional tasks [14, 28, 40].

Present work. We introduce RADT[1] (Reach-Avoid Decision Transformer), a reward-free, data-driven, offline RL approach for goal-conditioned and avoid-region-conditioned reach-avoid learning (Figure 1). RADT is a decision transformer that represents goals and avoid-regions as prompt tokens. This formulation decouples the reach-avoid specification from the state representation and enables zero-shot generalization to arbitrary numbers and sizes of avoid-regions. RADT learns policies entirely from random-policy trajectories using a novel combination of goal and avoid-region hindsight relabeling, with no need for reward/cost functions. Our main contributions include: ① A prompting framework for reach-avoid learning that encodes goals and avoid-regions as discrete prompt tokens, allowing flexible and interpretable conditioning of behavior at test time. ② A novel avoid-region hindsight relabeling strategy that allows the model to learn successful avoid behavior from suboptimal data in a fully data-driven manner, without the need for reward/cost functions. ③ A decision transformer model trained on random-policy data using this approach, with benchmarking across robotics and biological domains evaluates generalization to out-of-distribution avoid-region sizes and counts. ④ Strong results showing that RADT generalizes zero-shot to 12 unseen reach-avoid configurations, outperforming 3 existing methods retrained directly on those configurations.

2 Desiderata/Desirable Properties of Reach-Avoid RL Models

Notation. We first establish the notation used throughout this work. A trajectory τ of length T is a sequence of alternating states and actions collected by an agent making sequential decisions in an environment:

$$\tau = (\mathbf{s}_1, \mathbf{a}_1, \mathbf{s}_2, \mathbf{a}_2, \ldots, \mathbf{s}_T, \mathbf{a}_T)$$

where $\mathbf{s}_t \in \mathcal{S}$ is the state of the agent at timestep t and $\mathbf{a}_t \in \mathcal{A}$ is the action the agent takes from $\mathbf{s}_t$. $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ represents the state space and $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ represents the action space, which are the spaces

[1]RADT code: https://github.com/mims-harvard/RADT

of all possible states and actions, respectively. All RL models in this work learn a deterministic policy $\pi(\mathbf{a}_t \mid \cdot)$ that selects the most preferred action $\hat{\mathbf{a}}_t$ for the agent to take at timestep t given contextual inputs, typically including at least the current state $\mathbf{s}_t$ (i.e., $\pi(\mathbf{a}_t \mid \mathbf{s}_t)$). At each timestep of a trajectory, a *state transition* occurs, where the agent takes an action $\mathbf{a}$ from their current state $\mathbf{s}$ and ends up in a new state $\mathbf{s}'$. This is denoted by the transition tuple $(\mathbf{s}, \mathbf{a}, \mathbf{s}')$. If used, the reward function $r(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ and cost function $\texttt{cost}(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ return scalar values for each transition tuple in the trajectory. When applicable, these values are included in the trajectory as follows:

$$\tau = (\mathbf{s}_1, \mathbf{a}_1, r_1, c_1, \mathbf{s}_2, \mathbf{a}_2, \dots)$$

In goal-conditioned settings, a goal state $\mathbf{g} \in \mathcal{S}$ is provided as an additional input, yielding conditional functions such as $\pi(\mathbf{a} \mid \cdot, \mathbf{g})$ or $r(\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{g})$. In avoid-region-conditioned settings with n_{avoid} avoid-regions $\mathbf{b}_j : j \in \{1, 2, ..., n_{\text{avoid}}\}$, similar conditioning applies. We refer to the center of avoid-region $\mathbf{b}_j$ as its avoid centroid, denoted $\texttt{centroid}(\mathbf{b}_j)$.

Desiderata/Desirable properties. Reach-avoid problems introduce a dual behavioral objective: the agent must reach a desired target state while avoiding explicitly defined regions of the state space. Reach-avoid models that satisfy this behavioral objective under real-world deployment constraints need to achieve the following key properties:

Property 1 (P1): Pre-collected offline datasets with no online fine-tuning. The model must learn solely from offline datasets $\mathcal{D}$ containing of pre-collected trajectories $\tau^{(i)}$, with no reliance on online fine-tuning. In safety-critical applications, online exploration may be infeasible, especially when entering avoid-regions could cause irreversible harm or failure.

Property 2 (P2): Suboptimality-tolerant learning. The model must learn strong policies from offline datasets that contain only suboptimal trajectories, i.e., those that do not reach the goal or that violate the avoid constraint. Specifically, it should support super-demonstration performance by learning from trajectories $\tau^{(i)}$ that: **(2.1)** fail to reach the target goal state $\mathbf{g}$ at rollout time, and/or **(2.2)** pass through avoid-regions $\mathbf{b}_j$ rather than successfully avoiding them.

Property 3 (P3): Goal-conditioned generalization. The model must generalize to any arbitrarily specified goal state $\mathbf{g}$ at evaluation time, without additional training. In real-world scenarios, such as autonomous navigation or therapeutic reprogramming, the target goal is often specified dynamically and cannot be hardcoded at training time.

Property 4 (P4): Avoid-region-conditioned generalization. The model must be able to learn a policy that can avoid any dynamically specified avoid-region(s) $\mathbf{b}_j$ of the state space at evaluation time, without additional training/finetuning. This includes supporting changes in: **(4.1)** the number of avoid-regions n_{avoid}, **(4.2)** their locations $\texttt{centroid}(\mathbf{b}_j)$, and **(4.3)** their sizes (i.e., spatial extent of the state space around each $\texttt{centroid}(\mathbf{b}_j)$ to avoid).

Property 5 (P5): Reward-free learning. The model must learn reach-avoid behavior without requiring a manually designed reward or cost function. Reward shaping is often brittle and requires expert domain knowledge [12, 24, 25, 1], especially when preferences over reaching and avoiding are difficult to encode or conflict. Instead, one should be possible to specify goals and avoid-regions directly as inputs to the model. Additionally, reward-based safe RL methods are usually coupled to a constraint satisfaction requirement with respect to a cost function [2, 3, 62, 6], implying a priority hierarchy between goal-reaching and avoidance objectives. Such a priority hierarchy may be relevant for certain safety-critical environments, but does not apply to all reach-avoid problems (Section 1) and can even be actively incompatible in situations where avoid-regions are unavoidable (e.g., Section 5). Removing reward functions and cost functions eliminates the need to frame the problem as a constrained optimization problem and dissolves this implicit priority hierarchy. This relates to our choice of a Pareto-style evaluation framework discussed in Section 5 under "Evaluation Framework" and Appendix E under "Pareto vs. Constraint Optimization Evaluation Framework."

While many prior approaches address subsets of these properties, none satisfy all five simultaneously (Figure 1b). We discuss these limitations in detail in Section 3.

3 Related Work

We review four key areas in reach-avoid learning: offline goal-conditioned RL (OGCRL), offline safe RL (OSRL), offline goal-conditioned safe RL (OGCSRL), and decision transformer (DT) models. Figure 1b summarizes which properties each class of methods satisfies. Additional discussion appears in Appendix E.

Offline Goal-Conditioned RL. OGCRL methods aim to learn policies that generalize to arbitrary goals specified at evaluation time, typically by conditioning on goal states and applying techniques such as hindsight goal relabeling [4]. **Reward-based OGCRL** methods [65, 26, 38] optimize a policy $\pi(\mathbf{a}|\mathbf{s}, \mathbf{g})$ to maximize a goal-conditioned reward function $r(\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{g})$. While this framework supports goal generalization (P3), it does not satisfy P5, as it requires designing reward functions that capture both goal-reaching and avoid desires, which are difficult to construct in practice [24, 25, 12]. **Reward-free OGCRL** methods [10, 44, 15, 64, 37, 22] avoid altogether reward functions by learning directly from hindsight-relabeled trajectories directly using supervised learning. This more modern wave of data-driven methods satisfy P5 but do not inherently support avoid-region conditioning (P4), as they cannot incorporate constraints beyond the goal.

Offline Safe RL. OSRL methods [68, 30, 62, 29, 32] are designed for safety-critical tasks, learning policies that satisfy constraints specified via cost functions. A typical approach defines a cost function $\texttt{cost}(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ and learns a policy that maximizes reward return $\sum_t r_t$ subject to a cost return constraint $\sum_t c_t < k$, often via Lagrangian relaxation [52]. These methods can enforce avoid behavior, but they are not generally goal-conditioned (P3), and do not support dynamic conditioning on varying avoid-region configurations (P4). In addition, these methods fail P5 due to their reliance on handcrafted reward/cost functions and constrained optimization setups with inflexible inherent priority hierarchies.

Offline Goal-Conditioned Safe RL. This hybrid category combines elements of OGCRL and OSRL and comes closest to satisfying the full set of reach-avoid properties. Representative methods include Recovery-based Supervised Learning (RbSL) and Actionable Models with Lagrangian Constraints (AM-Lag) [6, 7, 52]. These models construct an augmented state space $\mathcal{S}^+ \subseteq \mathbb{R}^{d_s+d_s+n_{\text{avoid}} \cdot d_s}$ containing the agent's state $\mathbf{s} \in \mathbb{R}^{d_s}$, goal state $\mathbf{g} \in \mathbb{R}^{d_s}$, and n_{avoid} avoid centroids $\texttt{centroid}(\mathbf{b}_j) \in \mathbb{R}^{d_s}$, then learn policies with goal-conditioned and avoid-conditioned objectives. This design satisfies P3 and P4.2, enabling generalization to arbitrary $\mathbf{g}$ and avoid-region locations. However, these methods do not satisfy P4.1, as the number of avoid-regions n_{avoid} is fixed in the state space dimension, requiring retraining to accommodate more regions. They also fail to support P4.3, as the spatial extent of avoid-regions is encoded only in the cost function, requiring redefinition and retraining when region size changes. Furthermore, although these models could in principle satisfy P2.2 (learning from trajectories that violate avoid-regions), the training data in [6] is collected from environments with impassable obstacles, meaning no training trajectories actually pass through avoid-regions (Figure 3b), failing to demonstrate robustness to this type of data suboptimality. See Appendix C.1 for details.

Decision Transformers. DT models [8, 22, 67, 60, 58, 36, 33] represent a class of offline RL approaches that frame policy learning as sequence modeling. DTs use causal transformer architectures that take as input a trajectory τ and autoregressively predict actions, $\pi(\mathbf{a}_t|\tau_{1:t-1}, \mathbf{s}_t)$. Prompting has recently been introduced as a mechanism to extend DTs to goal-conditioned settings [63, 66]. We build our method off of MGPO [66], which introduces prompt-based conditioning on arbitrary goals, enabling zero-shot generalization across goal states, but not avoid-regions.

4 Reach-Avoid Decision Transformer (RADT)

In this section, we describe the main components of our method, RADT (Figure 2). Similar to MGPO, our method is based on a causal Transformer architecture and utilizes prompts to specify goal states (satisfying P3). However, unlike MGPO, RADT additionally allows for the specification of avoid-regions in the prompt (satisfying P4) and does not require reward-driven online prompt optimization (satisfying P1, P2, P5).

Prompt Tokens and Avoid-Region Representation

RADT takes in a prompt p to be presented to the Transformer model before a trajectory τ. RADT's autoregressive prediction of the next action during rollout is thus additionally conditioned on this prompt: $\pi(\mathbf{a}_t|p, \tau_{0:t-1}, \mathbf{s}_t)$. The prompt is structured as follows (Figure 2a):

$$p = (z, i_b, \mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_{n_{\text{avoid}}}, i_g, \mathbf{g}, e)$$

Note that there are six different types of tokens present in the prompt:

① **The avoid success indicator (SI) token,** $z : z \in \{0, 1\}$, indicates whether the trajectory τ following the prompt successfully circumvents the avoid state: $z = 1$ if all states in τ exist outside

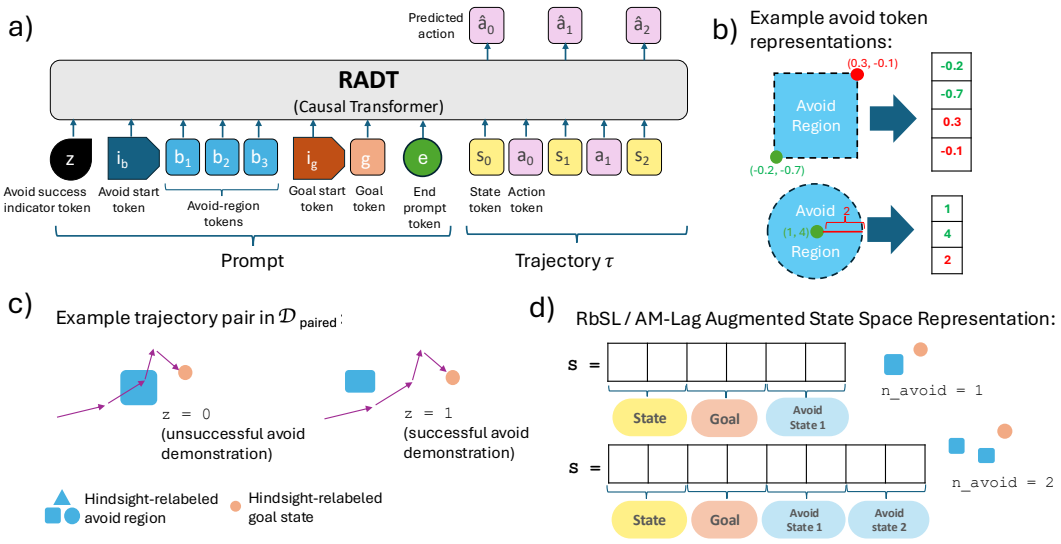


Figure 2: (a) RADT receives goal states and avoid-regions as prompt inputs. (b) avoid-regions are represented as avoid-region tokens in the prompt, which can be any vector representation that contains position and size information. The box and spherical representations shown here are examples of valid representations. (c) For each offline trajectory, we generate two versions: one that violates a sampled avoid-region and one that avoids it. Both are labeled with an avoid success token z. (d) Prior models encode avoid-regions via augmented state vectors, which grow with the number of avoid-regions, preventing zero-shot generalization to unseen avoid counts.

of all avoid-regions $\mathbf{b}_j : j \in \{1, 2, ..., n_{\text{avoid}}\}$ and $z = 0$ otherwise. This allows us to train the model on trajectories that demonstrate both successful and unsuccessful avoid behaviors; this is important for the model to explicitly learn what *not* to do (see Section 5, Results 4). During evaluation time, we *always* condition on $z = 1$ to achieve optimal avoid behavior. ② **The avoid start (AS) token**, i_b, explicitly indicates to the model that the upcoming tokens represent undesirable avoid-regions. This is to clearly distinguish the avoid-region tokens from the SI token. ③ **The avoid-region tokens**, $\mathbf{b}_j \in \mathcal{S} \subseteq \mathbb{R}^{d_a} : j \in \{1, 2, ..., n_{\text{avoid}}\}$, represents the n_{avoid} avoid-regions in the state space we would like the agent to circumvent. Our approach is not constrained to any particular vector representation of avoid-regions, as long as the representation contains information about both the position (the centroid) and the spatial extent of the avoid-region. Detailed overview of the representations we use can be found in Appendix A.2. Since prompts can consist of any arbitrary number of avoid-region tokens $\mathbf{b}_j$ and the avoid-region tokens can represent regions of any arbitrary size at evaluation time, this satisfies P4.1, P4.2, and P4.3. ④ **The goal start (GS) token**, i_g, explicitly indicates to the model that the next token to be provided represents a desirable goal token. This is to clearly distinguish the goal token in the prompt from the avoid-region tokens. ⑤ **The goal token**, $\mathbf{g} \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$, represents the goal state we would like the agent to achieve. This can be set to any state in the state space at evaluation time (satisfying P3). ⑥ **The prompt end token**, e, explicitly marks the end of the prompt, indicating that the next token marks the beginning of the main input sequence, τ.

The initial representations of tokens of different types have different dimensionalities, as described above, but they are all projected into a shared latent embedding space to acquire common dimensionality before being passed through the rest of the transformer architecture. See Appendix B.1 for details regarding how these different token types are embedded. Since we use prompts to specify the goal and avoid-region information, we do not need to work with an augmented state space $\mathcal{S}^+$, unlike RbSL and AM-Lag (see Section 3), providing us with greater zero-shot flexibility.

Avoid-Region Relabeling and Training Data

We specifically consider the scenario in which the training dataset $\mathcal{D}$ contains training trajectories τ that are generated from a purely random policy (satisfying P1 and P2). For each training trajectory $\tau \in \mathcal{D}$, we relabel the last state $\mathbf{s}_T$ as the goal state $\mathbf{g}$ in hindsight. Additionally, we can also relabel *avoid-regions* in hindsight, a novel strategy that gets rid of the need to use a cost function to learn desirable avoid behavior. The intuition for hindsight avoid-region relabeling (HAR) is similar to goal relabeling: it does not matter whether the policy that collected the training trajectory was actually

intending to circumvent the hindsight-relabeled avoid-region; we can still treat it as if the region was "meant" to be avoided, as the trajectory demonstrates how to *not* pass through that region.

The details of our HAR approach are described in Appendix A.3. HAR produces a dataset $\mathcal{D}_{\text{paired}}$ that contains a pair of trajectories $(\tau_{\text{orig}}, \tau_{\text{copy}})$ for each trajectory τ in the original dataset $\mathcal{D}$. One trajectory in each pair contains relabeled avoid regions that the trajectory successfully avoids and receives the label $z = 1$. The other contains avoid regions that the trajectory violates and receives the label $z = 0$. The two trajectories in a pair share the same states and actions, so the label z separates avoidance success and failure from differences between the trajectories (Figure 2c). We observe that this pairing enables credit assignment to behaviors that lead to avoidance violations, and the model learns to reduce these behaviors (Section 5, Results 4). At evaluation/deployment, we condition the model on $z = 1$. We format trajectories in $\mathcal{D}_{\text{paired}}$ with the prompt structure from the section above and train a causal transformer with a causal language modeling objective. Appendix B provides the model and training specifications.

5 Experiments

We evaluate RADT in three sets of reach-avoid environments: `FetchReachObstacle`, `MazeObstacle`, and `Cardiogenesis`. The first two are adapted from Gymnasium Robotics tasks [9], but with added avoid-regions. In these environments, we compare RADT against two OGCSRL baselines (RbSL [6] and AM-Lag [6, 7]) as well as Weighted Goal-Conditioned Supervised Learning (WGCSL) [64], a strong OGCRL method that does not explicitly account for avoid-regions. To evaluate domain generality, we also test RADT in a biological setting: cell reprogramming. Additionally, we perform a series of ablation and stress-testing experiments to validate the design decisions in our approach and evaluate how RADT reacts under more extreme out-of-distribution (OOD) avoid specifications, varying mixtures of random vs. expert data, and alternative representations for avoid-regions. Additional task/environment details are provided in Appendix C.

Evaluation framework. We treat reach-avoid problem as the Pareto optimization problem $\max_\theta \big(J_{\text{reach}}(\theta), J_{\text{avoid}}(\theta) \big)$ with two separate objectives to optimize for that may at times be at conflict with each other: (1) the goal-reaching ("reach") objective $J_{\text{reach}}(\theta) = \mathbb{E}_{\tau\sim\pi_\theta(\cdot|g,B)}\big[\, G(\tau, g) \,\big]$ and (2) the avoidance objective $J_{\text{avoid}}(\theta) = -\mathbb{E}_{\tau\sim\pi_\theta(\cdot|g,B)}\big[\, M(\tau, B) \,\big]$, where g is the goal, $B = [b_1, b_2 \dots]$ are the avoid-regions, G is the goal-reaching indicator function with leniency δ defined as $G(\tau, g) = \mathbb{1}\left[d(s_L, g) \leq \delta\right]$ where s_L is the last state in τ, and M is the MNC function defined as $M(\tau, B) = \frac{1}{|\tau|}\sum_{t=1}^{|\tau|} \mathbb{1}\left[s_t \in \bigcup_{j=1}^{n} b_j \right]$. While the two objectives could be combined into a single scalar objective and evaluation metric (e.g., the constrained optimization objectives used in many OSRL works [68, 30, 62, 29, 32]), that would involve predetermining and implying an optimal tradeoff between the two objectives, which is subjective to different scenarios and user preferences; therefore, we do not inherently define the optimal tradeoff between the two objectives and leave define both objectives defined and evaluated separately. Thus, if one method excels on one objective but underperforms on another objective compared to a second method, there is no objective "better method" in such a case. However, if one method were to dominate on one objective without being significantly worse on the other objective, it is said the method is *Pareto dominant*. Under this Pareto optimization framework, the following results show that RADT is Pareto dominant to baselines on the majority of experimental settings, and is incomparable to the other methods on the rest (*but is never Pareto-dominated by a baseline*). Further discussion in included in Appendix E under "Pareto vs. Constrained Optimization Evaluation Framework."

Results 1: Fetch Reach Environment and Generalization to Varying Avoid-Region Sizes

Environment. The `FetchReachObstacle` environment requires a robotic arm to reach a goal location while avoiding a randomly positioned box, treated as the avoid-region under our formulation (Figure 3a). The positions of the end-effector, goal, and avoid-box are randomly sampled each episode. Unlike prior work such as [6], the robot can pass through the avoid-box, allowing us to isolate avoid-region reasoning without hard state space constraints (Figure 3b).

Data. We collect $2 * 10^6$ timesteps of random-policy trajectories. Goal relabeling and HAR is then applied, using a contour-based sampling strategy for diverse box placements (Appendix A.3.2).

Evaluation metrics. While RADT is trained without a cost function, we define one for evaluation: $\texttt{cost}(s_t, a_t, s_{t+1}) = 1$ if s_{t+1} is inside an avoid-box and 0 otherwise. We evaluate models using the mean normalized cost return (MNC), computed as the average per-step cost:

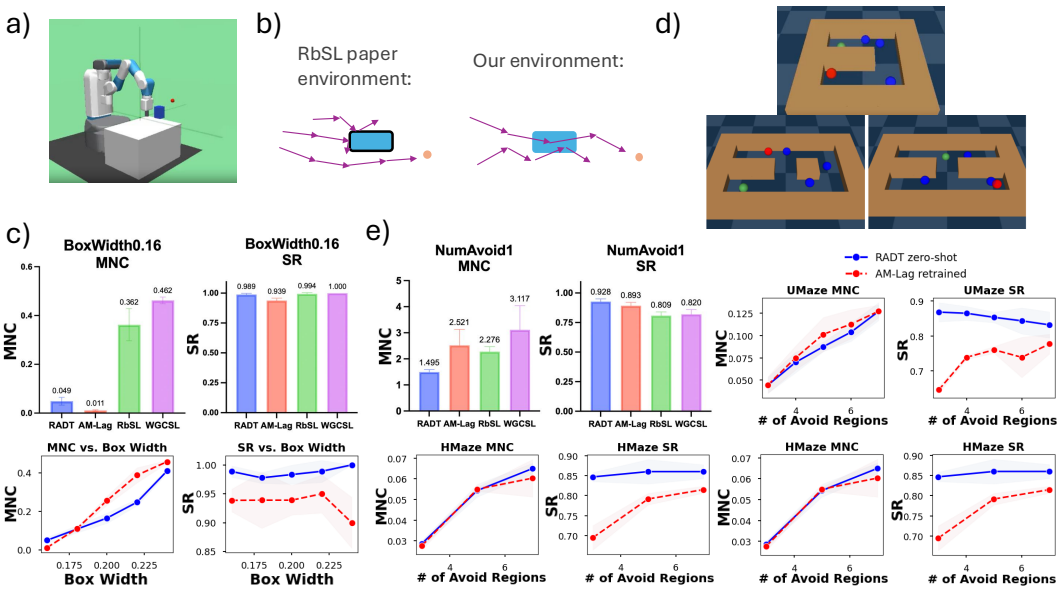


Figure 3: (a) Visualization of the `FetchReachObstacle` environment. The red point is the goal; the blue box is the avoid-region. (b) Unlike prior setups, the robot arm can pass through avoid-boxes, allowing training data to include violations so that we can test the models' abilities to learn from suboptimal data. (c) RADT and AM-Lag achieve state-of-the-art MNC and SR on in-distribution box sizes and generalizes zero-shot to OOD avoid-box sizes, matching or surpassing the best baseline (AM-Lag), which needs to be *retrained* on every new avoid-box size (i.e., not zero-shot). (d) Visualization of the `UMazeObstacle`, `AMazeObstacle`, and `HMazeObstacle` environments, with red goal, blue avoid-regions, and green agent. (e) RADT outperforms all baselines on MNC and SR in the in-distribution single-avoid setting for `UMazeObstacle` and generalizes zero-shot to OOD numbers of avoid-regions on all maze tasks, matching the best `retrained` baseline (AM-Lag) in MNC and surpassing it in SR. Error bars show ±1 standard deviation.

$\text{MNC}(\tau) = \frac{1}{|\tau|} \sum_{(s,a,s') \in \tau} \texttt{cost}(s, a, s')$. We also report the goal-reaching success rate (SR), defined as the proportion of episodes that reach the goal. Evaluation is performed over 60 episodes.

Setup. We train all models using environments with a fixed avoid-box width (or, in the case of RADT, a max hindsight-relabeled box width) of 0.16. RbSL, AM-Lag, and WGCSL are trained using the above cost function. We then evaluate all models in environments in-distribution (ID) with the same 0.16 avoid-box width. To assess generalization, we evaluate RADT zero-shot on avoid-boxes with out-of-distribution (OOD) widths ranging from 0.16 to 0.24 (1.5× larger). In contrast, baseline models do not support zero-shot generalization to new avoid-box sizes and must be retrained on separate offline datasets generated for each new size (Sections 3 and 4, and Appendix C.1). This creates a disadvantageous setting for RADT, which is evaluated without retraining, while baselines are retrained for each test condition.

Results. Results are shown in Table 1. For the ID case (box width 0.16), RADT performs comparably to AM-Lag in MNC and better than RbSL and WGCSL. For SR, RADT outperforms AM-Lag and matches or exceeds the other baselines (Figure 3c). This confirms that both RADT and AM-Lag are competitive for reach-avoid learning, with RADT slightly favoring goal-reaching and AM-Lag slightly favoring cost minimization. In the OOD setting, where avoid-box sizes exceed those seen during training, RADT continues to perform strongly despite being evaluated zero-shot. Across all OOD widths, RADT matches or exceeds the MNC of retrained AM-Lag models and substantially outperforms retrained RbSL and WGCSL models. It also maintains comparable SR to the retrained baselines (Figure 3c). These results show that RADT generalizes to unseen avoid-region sizes, even outperforming methods that are *retrained* for each evaluation setting.

Results 2: Maze Environments and Generalization to Varying Numbers of Avoid-Regions

Environment. The `MazeObstacle` environments require a point agent to navigate through a maze to reach a randomly sampled goal location while avoiding randomly placed circular obstacles. As in `FetchReachObstacle`, these obstacles are soft constraints that the agent may pass through, and we refer to them as "avoid-regions." Unlike in `FetchReachObstacle`, the maze itself imposes hard constraints, introducing impassable regions of the state space. This tests RADT's ability to reason over both user-specified avoid-regions and inherent environmental constraints. We evaluate RADT

Table 1: Results for `FetchReachObstacle` with varying box sizes (avg. over 3 seeds).

Box Width	RADT$^{\#}$		AM-Lag		RbSL		WGCSL	
	MNC	SR	MNC	SR	MNC	SR	MNC	SR
0.16	$\mathbf{0.049}_{\pm 0.016}$	$\mathbf{0.989}_{\pm 0.01}$	$\mathbf{0.011}_{\pm 0.003}$	$0.95_{\pm 0.029}$	$0.362_{\pm 0.066}$	$\mathbf{0.994}_{\pm 0.01}$	$0.462_{\pm 0.0148}$	$\mathbf{1.0}_{\pm 0.0}$
0.18	$\mathbf{0.107}_{\pm 0.018}$	$0.978_{\pm 0.009}$	$\mathbf{0.11}_{\pm 0.01}$	$0.95_{\pm 0.033}$	$0.484_{\pm 0.016}$	$\mathbf{1.0}_{\pm 0}$	$0.513_{\pm 0.048}$	$\mathbf{0.994}_{\pm 0.01}$
0.20	$\mathbf{0.164}_{\pm 0.027}$	$\mathbf{0.983}_{\pm 0.017}$	$0.255_{\pm 0.023}$	$0.955_{\pm 0.025}$	$0.571_{\pm 0.013}$	$\mathbf{1.0}_{\pm 0.0}$	$0.588_{\pm 0.066}$	$\mathbf{1.0}_{\pm 0.0}$
0.22	$\mathbf{0.247}_{\pm 0.018}$	$\mathbf{0.989}_{\pm 0.019}$	$0.387_{\pm 0.039}$	$0.944_{\pm 0.02}$	$0.64_{\pm 0.053}$	$\mathbf{0.99}_{\pm 0.01}$	$0.699_{\pm 0.019}$	$\mathbf{0.994}_{\pm 0.01}$
0.24	$\mathbf{0.409}_{\pm 0.012}$	$\mathbf{1.0}_{\pm 0.0}$	$\mathbf{0.457}_{\pm 0.008}$	$0.964_{\pm 0.017}$	$0.701_{\pm 0.019}$	$\mathbf{1.0}_{\pm 0}$	$0.729_{\pm 0.038}$	$\mathbf{0.989}_{\pm 0.01}$

$^{\#}$ = Model capable of zero-shot generalization. Results highlighted in blue are zero-shot results.

Table 2: Results for `MazeObstacle` with varying number of avoid states (avg. over 3 seeds).

# Avoid	UMazeObstacle							
	RADT$^{\#}$		AM-Lag		RbSL		WGCSL	
	MNC (1e-2)	SR	MNC (1e-2)	SR	MNC (1e-2)	SR	MNC (1e-2)	SR
1	$\mathbf{1.495}_{\pm 0.096}$	$\mathbf{.928}_{\pm 0.022}$	$2.521_{\pm 0.613}$	$0.893_{\pm 0.029}$	$2.276_{\pm 0.193}$	$0.809_{\pm 0.031}$	$3.117_{\pm 0.922}$	$0.82_{\pm 0.041}$
3	$\mathbf{4.455}_{\pm 0.895}$	$\mathbf{0.868}_{\pm 0.028}$	$\mathbf{4.47}_{\pm 0.94}$	$0.645_{\pm 0.01}$	$5.857_{\pm 0.754}$	$0.175_{\pm 0.054}$	$6.92_{\pm 0.404}$	$\mathbf{0.842}_{\pm 0.033}$
4	$\mathbf{7.006}_{\pm 1.156}$	$\mathbf{0.865}_{\pm 0.02}$	$7.511_{\pm 1.342}$	$0.738_{\pm 0.006}$	$7.648_{\pm 0.357}$	$0.768_{\pm 0.043}$	$9.62_{\pm 1.701}$	$\mathbf{0.852}_{\pm 0.043}$
5	$\mathbf{8.75}_{\pm 0.531}$	$\mathbf{0.853}_{\pm 0.015}$	$10.14_{\pm 1.9}$	$0.76_{\pm 0.01}$	$9.622_{\pm 0.531}$	$0.053_{\pm 0.012}$	$10.28_{\pm 0.503}$	$\mathbf{0.807}_{\pm 0.033}$
6	$\mathbf{10.38}_{\pm 1.015}$	$\mathbf{0.843}_{\pm 0.038}$	$11.278_{\pm 1.629}$	$0.738_{\pm 0.058}$	$11.35_{\pm 1.604}$	$0.755_{\pm 0.065}$	$12.364_{\pm 1.414}$	$\mathbf{0.9}_{\pm 0.017}$
7	$\mathbf{12.7}_{\pm 1.0}$	$\mathbf{0.832}_{\pm 0.038}$	$\mathbf{12.72}_{\pm 0.702}$	$0.777_{\pm 0.028}$	$24.17_{\pm 3.65}$	$0.002_{\pm 0.003}$	$14.48_{\pm 1.439}$	$\mathbf{0.825}_{\pm 0.018}$

# Avoid	AMazeObstacle				HMazeObstacle			
	RADT$^{\#}$		AM-Lag		RADT$^{\#}$		AM-Lag	
	MNC (1e-2)	SR	MNC (1e-2)	SR	MNC (1e-2)	SR	MNC (1e-2)	SR
3	$\mathbf{3.1}_{\pm 0.361}$	$\mathbf{0.926}_{\pm 0.006}$	$4.467_{\pm 0.252}$	$0.805_{\pm 0.02}$	$\mathbf{2.867}_{\pm 0.153}$	$\mathbf{0.847}_{\pm 0.012}$	$\mathbf{2.767}_{\pm 0.208}$	$0.695_{\pm 0.03}$
5	$\mathbf{4.933}_{\pm 0.252}$	$\mathbf{0.908}_{\pm 0.024}$	$5.473_{\pm 1.0}$	$0.808_{\pm 0.056}$	$\mathbf{5.433}_{\pm 0.306}$	$\mathbf{0.86}_{\pm 0.03}$	$\mathbf{5.5}_{\pm 0.1}$	$0.792_{\pm 0.014}$
7	$\mathbf{7.0}_{\pm 0.127}$	$\mathbf{0.922}_{\pm 0.013}$	$\mathbf{6.933}_{\pm 1.415}$	$0.86_{\pm 0.022}$	$6.5_{\pm 0.265}$	$\mathbf{0.86}_{\pm 0.017}$	$\mathbf{6.03}_{\pm 0.907}$	$0.815_{\pm 0.009}$

on three environments of varying difficulty: `UMazeObstacle`, `AMazeObstacle` (2 dead ends), and `HMazeObstacle` (4 dead ends) (Figure 3d). The Data/Evaluation pipelines are similar to those of `FetchReachObstacle`, with details in Appendix C.1.

Setup. We begin by training all models using data generated in environments with a single avoid-region, and evaluate in matching single-avoid-region settings (ID). To evaluate generalization, we then train and evaluate all models in environments with three avoid-regions, followed by testing in environments with 4-7 avoid-regions to assess zero-shot OOD performance. For each new number of avoid-regions, baseline models (RbSL, AM-Lag, WGCSL) are retrained with an appropriately expanded state space (Figure 2d; see Sections 3 and 4). In contrast, RADT is evaluated zero-shot without any retraining. This design intentionally favors the baselines, as they are tailored to each test setting, whereas RADT is held fixed across all configurations. Only the best-performing baseline in `UMazeObstacle` (AM-Lag) is evaluated on `AMazeObstacle` and `HMazeObstacle`.

Results. Results are shown in Table 2. Across all settings of `UMazeObstacle` (both ID and OOD), RADT achieves the lowest MNC, consistently outperforming all baselines. It also outperforms AM-Lag and RbSL on SR, and performs comparably or better than WGCSL. Importantly, for all evaluations with more than three avoid-regions, RADT is deployed zero-shot, while the baselines are retrained. Despite this disadvantage, RADT matches or exceeds their performance, demonstrating its ability to generalize to OOD numbers of avoid-regions without retraining.

In `AMazeObstacle` and `HMazeObstacle`, RADT outperforms AM-Lag in goal-reaching SR for every number of avoid regions. In MNC, RADT outperforms or matches AM-Lag, and its advantage decreases at 5 and 7 avoid regions, which are OOD for RADT and ID for AM-Lag.

Results 3: Applications in Biology: Zero-Shot Avoidance in Stochastic Cell Reprogramming

We apply RADT to a biomedical problem: cell reprogramming. The goal is to transition a cell from one gene expression state to another using sequences of genetic perturbations. This technique underpins regenerative medicine [57, 55], stem cell therapy [53, 18], and anti-aging strategies [41, 45]. However, intermediate gene expression states encountered during reprogramming may carry risks, e.g. tumorigenesis [61, 34]. Safe reprogramming fits the reach-avoid formulation: reach a target state while avoiding undesirable intermediate states (Figure 4a).

Environment. We introduce `Cardiogenesis`, an environment based on a well-established Boolean network model of gene expression dynamics during mouse cardiogenesis [20, 49]. This model is

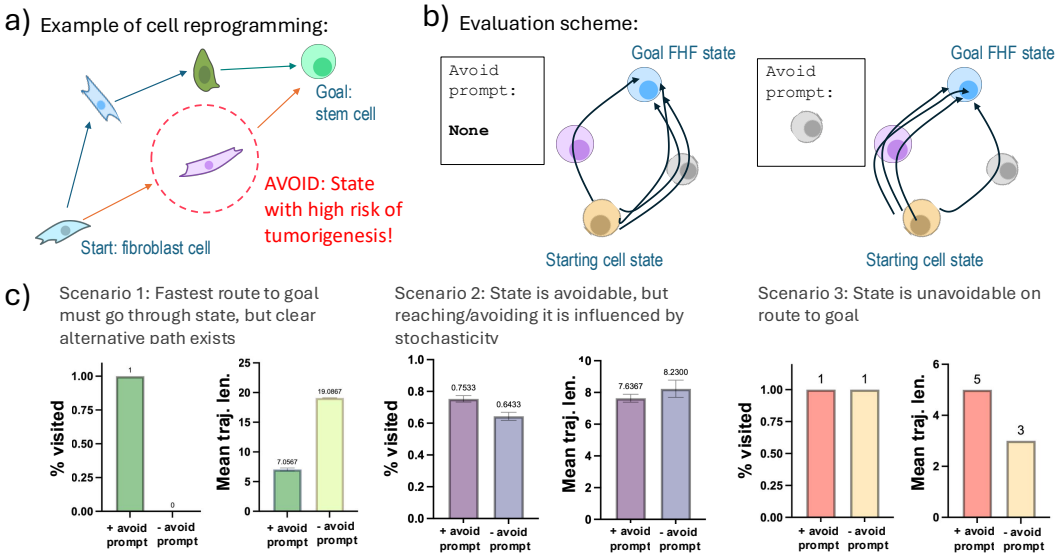


Figure 4: (a) Cell reprogramming involves sequential perturbations to reach a target gene expression state while avoiding undesirable intermediate states. (b) Evaluation pipeline: RADT is first run without an avoid-region token. The most frequently visited intermediate state is then added as an avoid-region token, and RADT is re-evaluated. (c) RADT reduces visitation rate to specified avoid states and minimizes time spent in unavoidable avoid states. Error bars show ±1 standard deviation. Some illustrations adapted from NIAID NIH BIOART (Appendix J).

described in detail in Appendix C.2. The key takeaway is that, unlike the Gymnasium Robotics environments, this domain features: (1) a fully discrete and combinatorial state-action space, (2) high-dimensional interdependencies between state variables due to Boolean logic, and (3) stochastic transitions. While RADT can be trivially adapted to this very different state-action structure, our primary baseline methods cannot, as we show in Appendix C.2 under "Baselines." See Appendix C.2 for full environment specifications.

Data. We generate $6 * 10^4$ random-policy timesteps of training data. Our HAR approach (Section 4) is directly applicable to this discrete setting with minimal modification (Appendix C.2).

Setup. We conduct a case study where the reprogramming goal is to reach the first heart field (FHF) state, a critical attractor in cardiac development [20, 49]. We train RADT on the offline data, then evaluate its ability to reach the FHF state starting from a distinct attractor state (Appendix C.2). At first pass, we run 200 evaluation episodes with no avoid-region specified. From these, we identify the most frequently visited intermediate state on the way to the goal. We then run a second set of 200 episodes, this time providing the most visited intermediate state as an *avoid-region token* in the prompt. We evaluate whether RADT is able to discover alternative reprogramming paths that avoid this state and compare both visitation rates and trajectory lengths. This setup is illustrated in Figure 4b. See Appendix C.2 for details.

Results. All evaluation trajectories successfully reach the FHF goal state, so we report only intermediate state visitation and trajectory lengths. Figure 4c summarizes results. We observe three distinct behavioral patterns, depending on the initial state. **Scenario 1.** The most visited intermediate (state A) is encountered in every trajectory when no avoid-region token is used. When A is included as an avoid-region token, RADT avoids it entirely and instead follows alternative, longer trajectories that always bypass A. **Scenario 2.** Without an avoid-region token, the most visited intermediate state (state B) appears in a large fraction of trajectories, and the remaining trajectories reach the goal state through alternative paths. State B is a frequent transition under the stochastic dynamics of the environment, and reaching the goal state does not depend on it. When we condition RADT on an avoid-region token for state B, the model reduces the visitation rate of state B in this noisy setting. **Scenario 3.** The most visited state (state C) is present in all trajectories regardless of prompt. When C is added as an avoid-region token, RADT cannot avoid passing through it, indicating that it is structurally unavoidable from the given initial state. However, RADT reduces the number of steps spent in C, shortening the portion of the trajectory that includes it. This suggests that even when avoidance is infeasible, the model learns to minimize time spent in unsafe states. These results highlight that RADT supports reach-avoid planning in discrete, structured, and stochastic domains, and

also exhibits flexible avoidance strategies, including temporal minimization of contact with undesirable states when full avoidance is not possible. Full results are in Appendix C.2.

Results 4: Additional Stress-Testing and Ablation Experiments

We conduct ablation studies to highlight the importance of the SI, AS, and GS prompt tokens and the usage of attention boosting to the prompt. We also perform additional stress-testing experiments to examine the extremities of RADT's OOD generalization, the empirical performance ceiling induced by the usage of random-policy data, and the RADT's robustness to changing avoid-region token representations. We summarize the results here, and leave the detailed experimental setup, results, and discussion in Appendix D.1,D.2,D.3,D.4 and Tables 6,7,9,10,12,13 in Appendix I.

Ablation experiments. We perform ablation experiments of the SI, AS, GS, and avoid-region tokens, avoid-region relabeling, and attention boosting. Performance drops with each of these ablations the importance of these elements to our methodological design. (Fig. 6, Tables 6,7,13).

Extremities of OOD generalization. To test the extremities RADT's zero-shot generalization, we test zero-shot performance up to 20 avoid-regions. We find that zero-shot RADT is still able to maintain comparable or superior MNC to the best *retrained* baseline at 10 avoid-regions ($>$300% of the maximum seen during training), demonstrating strong OOD generalization (Table 9). Notably, zero-shot RADT maintains a superior SR to retrained AM-Lag all the way up to 20 avoid-regions.

Effects of expert-policy data. To explore the performance ceiling induced by using purely random-policy data, we examine the effects of introducing varying amounts of expert-policy trajectories into the training data. We observe that the introduction of increasing amounts of expert-policy data does not significantly impact performance on ID box sizes, but does improve avoid ability (MNC) while *impairing* goal-reaching SR on OOD box sizes (Table 10). This suggests the particular expert trajectories used biases RADT to be more conservative in OOD avoid configurations.

Effects of changing avoid-region representation. We test RADT's flexibility with regards to handling various avoid-region token representations. We demonstrate that using an alternative representation for avoid-region tokens does not prevent RADT from learning strong reach-avoid policies, establishing that our approach is agnostic to avoid-region token representation (Table 12).

Simultaneous varying of multiple avoid-region axes. We show that RADT is capable of responding to variation in multiple axes of avoid-region specification simultaneously using a navigation environment that generates a random number (ranging from 2 to 4) of avoid-regions, each with random location, aspect ratio (from 1:1 to 1:2.5), and size (side lengths from 0.24 to 0.6). As shown in Table 14, RADT outperforms baselines in MNC and SR on this task.

Image-based observation spaces and avoid-region representations. To explore RADT's potential in handling 1) high-dimensional observation spaces and 2) unstructured, flexible avoid-region token representations that can capture abstract avoidance desires without knowledge of a pre-defined vectorization, we conduct a preliminary proof-of-concept demonstration of RADT applied to an image-based version of `UMazeObstacle` that we call `UMazeObstacleImage` (Fig. 12). In this environment, we present states, avoid-region tokens, and goal tokens as images encoded using a ResNet-18 model [19, 39]. RADT achieves superior MNC and comparable or superior goal-reaching SR to baselines in this environment, demonstrating its potential in being adapted to both high-dimensional tasks and flexible avoid-region representations that do not require precise knowledge of a strict, vectorized structure to specify avoid-regions (Table 15). Experiment setup is described in Appendix D.6.

6 Conclusion

We introduce RADT, a prompting-based offline RL approach for reach-avoid tasks that satisfies all key desiderata for flexible reach-avoid learning. RADT is trained entirely on suboptimal data using hindsight relabeling and generalizes zero-shot to unseen avoid-region configurations, achieving competitive or superior performance to SOTA methods retrained for each evaluation setting. RADT is also domain-agnostic, demonstrating strong performance in both robotics and cell biology domains. We attribute RADT's versatility and adaptability to its interpretable prompt-based design and fully data-driven, reward-free learning procedure. These properties enable RADT to serve as a general-purpose framework for sequential decision-making under dynamically shifting avoidance desires. Limitations of our model are described in Appendix G.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge support, in part, by NSF CAREER 2339524, Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists (ROADS) Program, Sanofi iDEA-iTECH Award, GlaxoSmithKline Award, Boehringer Ingelheim Award, Merck Award, Optum AI Research Collaboration Award, Pfizer Research, Gates Foundation (INV-079038), Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. This work was delivered as part of the Aurora project supported by the Cancer Grand Challenges partnership funded by Cancer Research UK (CGCAI1-Mar26/100001). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders. All authors declare no competing interests.

References

[1] Ahmed Abouelazm, Jonas Michel, and J Marius Zöllner. A review of reward functions for reinforcement learning in the context of autonomous driving. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, June 2024.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 22–31. JMLR.org, 2017.

[3] Eitan Altman. Constrained markov decision processes with total cost criteria: Occupation measures and primal LP. *Math. Methods Oper. Res. (Heidelb.)*, 43(1):45–72, February 1996.

[4] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. 2017.

[5] Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint, 2021.

[6] Chenyang Cao, Zichen Yan, Renhao Lu, Junbo Tan, and Xueqian Wang. Offline goal-conditioned reinforcement learning for safety-critical tasks with recovery policy. *arXiv preprint arXiv:2403.01734*, 2024.

[7] Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, and Sergey Levine. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

[8] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

[9] Rodrigo de Lazcano, Kallinteris Andreas, Jun Jet Tai, Seungjae Ryan Lee, and Jordan Terry. Gymnasium robotics, 2024.

[10] Benjamin Eysenbach, Tianjun Zhang, Ruslan Salakhutdinov, and Sergey Levine. Contrastive learning as goal-conditioned reinforcement learning. *arXiv preprint arXiv:2206.07568*, 2022.

[11] Meng Feng, Viraj Parimi, and Brian Williams. Safe multi-agent navigation guided by goal-conditioned safe reinforcement learning, 2025.

[12] Kilian Freitag, Kristian Ceder, Rita Laezza, Knut Åkesson, and Morteza Haghir Chehreghani. Curriculum reinforcement learning for complex reward functions. 2024.

[13] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[14] Briti Gangopadhyay, Zhao Wang, Jia-Fong Yeh, and Shingo Takamatsu. Integrating domain knowledge for handling limited data in offline RL. 2024.

[15] Dibya Ghosh, Abhishek Gupta, Justin Fu, Ashwin Reddy, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals without reinforcement learning. *ArXiv*, abs/1912.06088, 2019.

[16] R L Graham. An efficient algorith for determining the convex hull of a finite planar set. *Inf. Process. Lett.*, 1(4):132–133, June 1972.

[17] Sven Gronauer. Bullet-safety-gym: A framework for constrained reinforcement learning. Technical report, mediaTUM, 2022.

[18] Jingyang Guan, Guan Wang, Jinlin Wang, Zhengyuan Zhang, Yao Fu, Lin Cheng, Gaofan Meng, Yulin Lyu, Jialiang Zhu, Yanqin Li, Yanglu Wang, Shijia Liuyang, Bei Liu, Zirun Yang, Huanjing He, Xinxing Zhong, Qijing Chen, Xu Zhang, Shicheng Sun, Weifeng Lai, Yan Shi, Lulu Liu, Lipeng Wang, Cheng Li, Shichun Lu, and Hongkui Deng. Chemical reprogramming of human somatic cells to pluripotent stem cells. *Nature*, 605(7909):325–331, May 2022.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Franziska Herrmann, Alexander Groß, Dao Zhou, Hans A Kestler, and Michael Kühl. A boolean model of the cardiac gene regulatory network determining first and second heart field identity. *PLoS One*, 7(10):e46798, October 2012.

[21] Kai-Chieh Hsu*, Vicenç Rubies-Royo*, Claire J. Tomlin, and Jaime F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[22] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021.

[23] Naoki Katsura and Federico Baldassarre. Cosine annealing with warmup for pytorch. https://github.com/katsura-jp/pytorch-cosine-annealing-with-warmup.git, 2021.

[24] W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023.

[25] W. Bradley Knox and James MacGlashan. How to specify reinforcement learning objectives. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

[26] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. 2021.

[27] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. 2019.

[28] Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2022.

[29] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3703–3712. PMLR, 09–15 Jun 2019.

[30] Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation. In *International Conference on Learning Representations*, 2022.

[31] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[32] Haohong Lin, Wenhao Ding, Zuxin Liu, Yaru Niu, Jiacheng Zhu, Yuming Niu, and Ding Zhao. Safety-aware causal representation for trustworthy offline reinforcement learning in autonomous driving. *IEEE Robotics and Automation Letters*, 9(5):4639–4646, 2024.

[33] Qian Lin, Bo Tang, Zifan Wu, Chao Yu, Shangqin Mao, Qianlong Xie, Xingxing Wang, and Dong Wang. Safe offline reinforcement learning with real-time budget constraints. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[34] Ying-Chu Lin, Cha-Chien Ku, Kenly Wuputra, Chung-Jung Liu, Deng-Chyang Wu, Maki Satou, Yukio Mitsui, Shigeo Saito, and Kazunari K Yokoyama. Possible strategies to reduce the tumorigenic risk of reprogrammed normal and cancer cells. *Int. J. Mol. Sci.*, 25(10):5177, May 2024.

[35] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, and Ding Zhao. Datasets and benchmarks for offline safe reinforcement learning. *Journal of Data-centric Machine Learning Research*, 2024.

[36] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[37] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. *Conference on Robot Learning (CoRL)*, 2019.

[38] Yecheng Jason Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned reinforcement learning via f-advantage regression. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[39] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery.

[40] Soichiro Nishimori, Xin-Qiang Cai, Johannes Ackermann, and Masashi Sugiyama. Offline reinforcement learning with domain-unlabeled data. 2024.

[41] Patrick T Paine, Ada Nguyen, and Alejandro Ocampo. Partial cellular reprogramming: A deep dive into an emerging rejuvenation technology. *Aging Cell*, 23(2):e14039, February 2024.

[42] J.-S Park and S.-J Oh. A new concave hull algorithm and concaveness measure for n-dimensional datasets. *Journal of Information Science and Engineering*, 29:379–392, 03 2013.

[43] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations (ICLR)*, 2025.

[44] Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[45] Beatriz Pereira, Francisca P Correia, Inês A Alves, Margarida Costa, Mariana Gameiro, Ana P Martins, and Jorge A Saraiva. Epigenetic reprogramming as a key to reverse ageing and increase longevity. *Ageing Res. Rev.*, 95(102204):102204, March 2024.

[46] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[47] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. 2019.

[48] Pedro Luiz Silva, Fadhel Ayed, Antonio De Domenico, and Ali Maatouk. Pay attention to what matters. In *MINT: Foundation Model Interventions*, 2024.

[49] Vivek Singh. Optimizing sequential gene expression modulation for cellular reprogramming - coupled boolean modeling and reinforcement learning based method. March 2024.

[50] Oswin So, Cheng Ge, and Chuchu Fan. Solving minimum-cost reach avoid using reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[51] Vlad Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models, 2025.

[52] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. 2020.

[53] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, August 2006.

[54] Zhixiong Tang. concave_hull. https://github.com/cubao/concave_hull.git, 2022.

[55] Lakshmy Vasan, Eunjee Park, Luke Ajay David, Taylor Fleming, and Carol Schuurmans. Direct neuronal reprogramming: Bridging the gap between basic science and clinical application. *Front. Cell Dev. Biol.*, 9:681087, July 2021.

[56] Phan Vinh and Nguyen Dung. *Context-Aware Systems and Applications. 11th EAI International Conference, ICCASA 2022 Vinh Long, Vietnam, October 27–28, 2022 Proceedings.* 04 2023.

[57] Yue Wan and Yan Ding. Strategies and mechanisms of neuronal reprogramming. *Brain Res. Bull.*, 199(110661):110661, July 2023.

[58] Ruhan Wang and Dongruo Zhou. Safe decision transformer with learning-based constraints. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[60] Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. Elastic decision transformer. 2023.

[61] Kenly Wuputra, Chia-Chen Ku, Deng-Chyang Wu, Ying-Chu Lin, Shigeo Saito, and Kazunari K Yokoyama. Prevention of tumor risk associated with the reprogramming of human pluripotent stem cells. *J. Exp. Clin. Cancer Res.*, 39(1):100, June 2020.

[62] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized q-learning for safe offline reinforcement learning. *Proc. Conf. AAAI Artif. Intell.*, 36(8):8753–8760, June 2022.

[63] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, B. Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *Thirty-ninth International Conference on Machine Learning*, 2022.

[64] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline RL. In *International Conference on Learning Representations*, 2022.

[65] Wenyan Yang, Huiling Wang, Dingding Cai, Joni Pajarinen, and Joni-Kristen Kämäräinen. Swapped goal-conditioned offline reinforcement learning. 2023.

[66] Haoqi Yuan, Yuhui Fu, Feiyang Xie, and Zongqing Lu. Pre-trained multi-goal transformers with prompt optimization for efficient online adaptation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[67] Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. *CoRR*, abs/2202.05607, 2022.

[68] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Eben Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *The Twelfth International Conference on Learning Representations*, 2024.

[69] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Forty-second International Conference on Machine Learning*, 2025.

Appendix

A Data Preparation

A.1 Datasets

The download link for the preprocessed datasets, as well as the setup and preprocessing code used to generate these datasets, will be released with the code repository for the project. See Appendix H.

A.2 Avoid Representation Details

avoid-regions are represented in our experiments as *box coordinates* (Fig 2b) by default, though any vector representation could work. The box coordinate vector of an avoid-region $\mathbf{b}_j \in \mathbb{R}^{2*d_s}$ represents a "box" in the state space to be avoided. It is defined such that the first d_s entries represent the lower bounds of each of the state space dimensions (the "lower left corner" of the box in a 3D analogy) and the second d_s entries represent the upper bounds of each of the state space dimensions (the "upper right corner"):

$$\mathbf{b}_j = [\underbrace{l_1, l_2, ..., l_{d_s}}_{\text{lower bounds}}, \overbrace{u_1, u_2 ..., u_{d_s}}^{\text{upper bounds}}]$$

The policy should avoid guiding the agent into the region of the state space bounded by this box. To see how we adapt this box representation for non-rectangular avoid-regions in the `MazeObstacle` (circular avoid-regions) and `Cardiogenesis` environments (discrete avoid states), refer to the "Avoid-Region Relabeling" sections of Appendix C.1 and Appendix C.2, respectively.

In addition to the box representation, we also conduct experiments using the following representation for spherical avoid-regions to demonstrate RADT's flexibility in handling different avoid-region token representations:

$$\mathbf{b}_j = [b_1, b_2, ..., b_{d_s}, r]$$

where $(b_1, b_2, ..., b_{d_s})$ represents the location of $\texttt{centroid}(\mathbf{b}_j)$ in the state space and r represents the radius of the avoid-region (Figure 2b). Results for the experiment are found in Appendix D.4.

A.3 Avoid Relabeling Details

A.3.1 Avoid Relabeling Two-Pass Algorithm

The two-pass hindsight avoid-region relabeling method introduced in Section 4 is described in more detail here and visualized in Algorithm 1 (first pass) and Algorithm 2 (second pass).

For each $\tau^{(i)} \in \mathcal{D}$, we carry out HAR in two passes. In the first pass, we randomly sample avoid-boxes $\mathbf{b}_j : j \in \{1, 2, ..., n_{\text{avoid}}\}$ of random sizes in the state space $\mathcal{S}$ and check whether any $s_t \in \tau^{(i)}$ violate any $\mathbf{b}_j : j \in \{1, 2, ..., n_{\text{avoid}}\}$. If there are no violations, then the SI for $\tau^{(i)}$ is set to $z^{(i)} = 1$, otherwise it is set to $z^{(i)} = 0$. In the second pass, we create a copy of the dataset, $\mathcal{D}_{\text{copy}}$, and go through the same process above with the trajectories ${\tau_{\text{copy}}}^{(i)} \in \mathcal{D}_{\text{copy}}$. This time, however, for a trajectory ${\tau_{\text{copy}}}^{(i)}$, we keep re-sampling avoid-boxes until the avoid success token ${z_{\text{copy}}}^{(i)}$ is the *opposite* of $z^{(i)}$ for the corresponding $\tau^{(i)}$ in the original $\mathcal{D}$.

Combining the datasets $\mathcal{D}$ and $\mathcal{D}_{\text{copy}}$ into $\mathcal{D}_{\text{paired}}$, we now have a *pair* of trajectories $(\tau^{(i)}_{\text{orig}}, \tau^{(i)}_{\text{copy}})$ for each $\tau^{(i)}$ in the original dataset, where one of $(\tau^{(i)}_{\text{orig}}, \tau^{(i)}_{\text{copy}})$ demonstrates successful avoid behavior and the other demonstrates unsuccessful avoid behavior.

A.3.2 More Sophisticated Avoid Centroid Sampling Strategies

The sampling of avoid centroids in Line 4 of Algorithm 1 can be done using naive uniform sampling over the state space (the default) or more sophisticated methods.

Contour-Based Sampling. One such sophisticated method we use samples avoid centroids that fit into the nooks of the broad contour of a training trajectory τ (Figure 5a). This is the sampling method used for the `FetchReachObstacle` environment (a very open environment with no inherent

restricted areas of the state space). We can acquire an outline of the general *contour* of a τ by calculating a concave hull of the data points $\mathbf{s}_t \in \tau$ in the state space $\mathcal{S}$ using the algorithm presented in [42] as implemented in the `concave_hull` library [54]. We denote the set of states in τ that belong to the convex hull as $\mathcal{S}_{\text{convex}}$. We can then find "nooks" (concave portions) of the trajectory by calculating the convex hull using the algorithm presented in [16] as implemented the `concave_hull` library, and then find the points that are part of the concave hull but not the convex hull; these points outline the concave nooks of the contour and we denote the set of these states as $\mathcal{S}_{\text{nook}}$. For each nook in trajectory τ, we then find the two convex hull points in $\mathcal{S}_{\text{convex}}$ bordering the set $\mathcal{S}_{\text{nook}}$ and sample a point between them as the avoid centroid. As a result, the trajectory of states *appears* to be "attempting" to avoid the hindsight-relabeled avoid centroid by wrapping around it (Figure 5a). Although the training trajectory was not trying to avoid this state, this does not matter because the resulting hindsight-relabeled trajectory does demonstrate how to circumvent an avoid state by wrapping around it. This sampling method is found to improve performance on the `FetchReachObstacle` environment (Figure 5b).

Sampling From a Limited Portion of the State Space. In certain state spaces, there may be regions where sampling an avoid centroid there would not provide helpful information. Thus, instead of sampling from the entire state space, we can just sample from just the portion of the state space $\mathcal{S}_{\text{limited}} \subset \mathcal{S}$ instead. This is done in the `MazeObstacle` and `Cariogenesis` environments, see Appendices C.1 and C.2 for details.

B Model/Training Details

B.1 Model Architectural Details and Hyperparameters

Representations and Embeddings for Prompt Tokens. All prompt tokens, regardless of type, are embedded into the same latent space as the action ($\mathbf{a}_t$) and state ($\mathbf{s}_t$) tokens in the main trajectory sequence: $\mathbb{R}^{d_h}$, where d_h is the hidden dimension/embedding dimension (corresponding to hyperparameter `embed_dim` below). **The avoid success indicator token** z is initially presented to the model as a one-hot vector: $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ if $z = 1$ and $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ if $z = 0$. This one-hot representation is then embedded into $\mathbb{R}^{d_h}$ via the learnable embedding matrix $E_z \in \mathbb{R}^{2\times d_h}$. **The avoid start token**, i_b, and **the goal start token**, i_g, are initially presented to the model as the one-hot vectors $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, respectively. They are then embedded into $\mathbb{R}^{d_h}$ via the learnable embedding matrix $E_i \in \mathbb{R}^{2\times d_h}$. The **prompt end start token**, e, is embedded into $\mathbb{R}^{d_h}$ as learnable vector $\mathbf{e} \in \mathbb{R}^{d_h}$. **Avoid-region tokens** $\mathbf{b}_j \in \mathbb{R}^{2d_s}$ and the **goal token** $\mathbf{g} \in \mathbb{R}^{d_s}$ are embedded into $\mathbb{R}^{d_h}$ via learnable embedding matrix $E_b \in \mathbb{R}^{2d_s\times d_h}$ and learnable embedding matrix $E_g \in \mathbb{R}^{d_s\times d_h}$, respectively.

Attention Boosting. We find that adding a bias `adelta` (a hyperparameter) to the attention logits to all prompt tokens, using the strategy described in [48], noticeably improves the instruction-following

Algorithm 1 Hindsight Avoid-Region Relabeling: Initial Pass

Require: state space $\mathcal{S}$, offline training dataset $\mathcal{D}$, maximum avoid-box width w_{max}, number of avoid states n_{avoid}

for training trajectory $\tau^{(i)}$ in $\mathcal{D}$ **do**
  Initialize avoid-regions list b_list$^{(i)}$ $\leftarrow$ []
  for j in 1, 2, ..., n_{avoid} **do**
    Randomly initialize avoid centroid $\mathbf{x}_j \in \mathcal{S}$
    Randomly choose avoid-box width $w \in [0, w_{\text{max}}]$
    avoid-box $\mathbf{b}_j \leftarrow$ concatenate($\mathbf{x}_j - \frac{w}{2}$, $\mathbf{x}_j + \frac{w}{2}$) ▷ box of width w centered around $\mathbf{x}_j$
    Append $\mathbf{b}_j$ to b_list$^{(i)}$
  end for
  if All none of the states in $\tau^{(i)}$ are in any of the avoid-boxes in b_list$^{(i)}$ **then**
    Avoid success $z^{(i)} \leftarrow 1$
  else
    Avoid success $z^{(i)} \leftarrow 0$
  end if
  Add $z^{(i)}$ and b_list$^{(i)}$ to the corresponding training prompt $p^{(i)}$ for $\tau^{(i)}$
end for

Algorithm 2 Hindsight Avoid-Region Relabeling: Second Pass

Require: state space $\mathcal{S}$, copy of the original offline training dataset $\mathcal{D}_{\text{copy}}$, maximum avoid-box width $w_{\max}$, number of avoid states n_{avoid}

for training trajectory ${\tau_{\text{copy}}}^{(i)}$ in $\mathcal{D}_{\text{copy}}$ **do**
  ${z_{\text{copy}}}^{(i)} \leftarrow z^{(i)}$
  while ${z_{\text{copy}}}^{(i)} = z^{(i)}$ **do**
    Initialize avoid-regions list b_list$^{(i)} \leftarrow []$
    for j in 1, 2, ..., n_{avoid} **do**
      Randomly initialize avoid centroid $\mathbf{x}_j \in \mathcal{S}$
      Randomly choose avoid-box width $w \in [0, w_{\max}]$
      avoid-box $\mathbf{b}_j \leftarrow$ concatenate($\mathbf{x}_j - \frac{w}{2}, \mathbf{x}_j + \frac{w}{2}$) ▷ box of width w centered around $\mathbf{x}_j$
      Append $\mathbf{b}_j$ to b_list$^{(i)}$
    end for
    if All none of the states in $\tau^{(i)}$ are in any of the avoid-boxes in b_list$^{(i)}$ **then**
      Avoid success ${z_{\text{copy}}}^{(i)} \leftarrow 1$
    else
      Avoid success ${z_{\text{copy}}}^{(i)} \leftarrow 0$
    end if
  end while
  Add ${z_{\text{copy}}}^{(i)}$ and b_list$^{(i)}$ to the corresponding training prompt ${p_{\text{copy}}}^{(i)}$ for ${\tau_{\text{copy}}}^{(i)}$
end for

ability of RADT to the prompt—i.e., improves goal-reaching behavior without negatively impacting avoid behavior (Figure 6).

Model Hyperparameters. The base causal transformer architecture we use is based on the HuggingFace implementation [59] of the GPT-2 architecture [46]; all model details that are not explicitly specified here are set to the default values for the HuggingFace implementation of GPT-2. We perform hyperparameter optimization (HPO) with a simple random search algorithm, using the RayTune framework [31]. The tunable model hyperparameters and their respective set of possible values in the search space are described in Table 3. Training hyperparameters are described in the next section.

Table 3: Model hyperparameters and search space

Hyperparameter	Description	Search space
`n_head`	Number of attention heads	`tune.choice([1, 3, 4, 6, 12])`
`n_layer`	Number of self-attention layers	`tune.choice([3, 6, 12])`
`embed_dim`	Size of latent embedding space (hidden dimension)	Fixed to be `64 * n_head`
`adelta`	Attention boosting bias to the prompt	`tune.choice([0, 1, 2, 3])`

The model hyperparameter configurations used are {`n_head=4, n_layer=4, embed_dim=256, adelta=2`} for the `FetchReachObstacle` environment, {`n_head=6, n_layer=6, embed_dim=384, adelta=1`} for the `MazeObstacle` environment, and {`n_head=6, n_layer=6, embed_dim=384, adelta=1`} for the `Cardiogenesis` environment.

B.2 Training Details

Loss Function. For a batch of size B of trajectories of length T, we use the following two-component loss function during training. The first component is the typical loss used in decision transformer models, the action loss $\mathcal{L}_{\text{action}}$. This is the mean squared error (MSE) between the predictions of next actions $\hat{\mathbf{a}}_\mathbf{t}$ based on the last-layer embeddings of $(p, \tau_{1:t-1}, \mathbf{s}_t)$ and the actual next actions $\mathbf{a}_\mathbf{t}$:

$$\mathcal{L}_{\text{action}} = \frac{1}{B} \sum_{i=1,2,\ldots,B} \frac{1}{T} \sum_{t=1,2,\ldots,T} (\widehat{\mathbf{a}_t^{(i)}} - \mathbf{a}_t^{(i)})^2$$

The second component is used to encourage RADT to learn how to be aware, at each timestep, whether or not the current state $\mathbf{s_t}$ violates any avoid-box $\mathbf{b}_j : j \in \{1, 2, ..., n_{\text{avoid}}\}$ in prompt p. We define the indicator k_t to indicate whether or not $\mathbf{s_t}$ violates any avoid-boxes $\mathbf{b}_j$ in the prompt, with $k_t = 1$ indicating that $\mathbf{s_t}$ does not violate any avoid-boxes and $k_t = 0$ indicating $\mathbf{s_t}$ violates at least one avoid-box. During training *only*, in addition to predicting the next action $\mathbf{a_t}$, we also make RADT predict k_t using the last-layer embeddings of $(p, \tau_{1:t-1}, \mathbf{s}_t)$. We define the box awareness loss $\mathcal{L}_{\text{avoid_awareness}}$ to be the binary cross entropy (BCE) loss between predicted $\hat{k_t}$ and actual k_t.

$$\mathcal{L}_{\text{avoid_awareness}} = \frac{1}{B} \sum_{i=1,2,...,B} \frac{1}{T} \sum_{t=1,2,...,T} k_t^{(i)} \log(\widehat{k_t^{(i)}}) + (1 - k_t^{(i)}) \log(1 - \widehat{k_t^{(i)}}))$$

The combined loss function is then:

$$\mathcal{L} = \mathcal{L}_{\text{action}} + \alpha \mathcal{L}_{\text{avoid_awareness}}$$

α is a tunable hyperparameter to balance the different components of the loss function. During hyperparameter optimization, we try values of α ranging from 0.1 to 10 (Fig. 9). Very large values of α (e.g., 5 and 10) de-prioritize $\mathcal{L}_{\text{action}}$ too much and significantly impairs learning a policy with strong goal-reaching (SR) and avoidance (MNC) behavior. While very small values of α (e.g., 0.1) results in the fastest convergence to a policy with strong SR and MNC performance, intermediate α values ($\alpha = 1$) allows the model to eventually converge on a policy with the lowest MNC, with equally good SR. Thus, we set it to 1.0 in all our experiments.

Stopping Criteria. We let all RADT models train for 10,000 training steps (no visual improvement in SR or MNC beyond that), checkpointing every 100 steps. As we use a similar training scheme as the one used in the original Prompt DT [63] and MGPO [66] papers, where we sample batches of trajectories with replacement, we cannot use "epoch" as a measurement of training progress. At every checkpointing iteration, we run an evaluation on the model in the same way we did in our experiments (described in Section 5). We choose the checkpoint with the best (lowest) MNC whose SR is within 0.05 from the checkpoint with the highest SR (to ensure we have a model that is not achieving seemingly good avoid ability by significantly sacrificing goal-reaching ability, e.g., not moving). We do the same process for all baselines, except we train the baseline for 500 epochs as is done in the RbSL paper [6] (no visual improvement in SR or MNC beyond that).

Training Hyperparameters. Training hyperparameters are listed with their respective search spaces in Table 4. The scheduler value `'lambdalr'` corresponds to the PyTorch scheduler `torch.optim.lr_scheduler.LambdaLR` and the scheduler value `'cosinewarmrestarts'` corresponds to the scheduler `CosineAnnealingWarmupRestarts` from the `cosine_annealing_warmup` library [23] library is used to tune training hyperparameters. Any hyperparameters not explicitly listed default to the values used in MGPO [66].

The chosen training hyperparameter configurations are {`batch_size=128, learning_rate=1e-4, scheduler='cosinewarmrestarts', T_0=1000, warmup_steps=500, alpha1=1`} for the `FetchReachObstacle` environment, {`batch_size=32, learning_rate=1e-4, scheduler='lamdalr', warmup_steps=1000, alpha1=1`} for the `MazeObstacle` environment, and {`batch_size=128, learning_rate=1e-4, scheduler='lamdalr', warmup_steps=1000, alpha1=1`} for the `Cardiogenesis` environment.

Table 4: Training hyperparameters and search space

Hyperparameter	Description	Search space
`batch_size`	Batch size	`tune.choice([32, 64, 128, 256])`
`learning_rate`	Maximum learning rate	`tune.choice([1e-4, 1e-5])`
`scheduler`	Learning rate scheduler	`tune.choice(['cosinewarmrestarts', 'lambdalr'])`
`warmup_steps`	Number of warmup steps	`tune.choice([500, 1000])`
`T_0`	`T_0` parameter to `CosineAnnealingWarmupRestarts`	Fixed to be `1000`
`T_mult`	`T_mult` parameter to `CosineAnnealingWarmupRestarts`	Fixed to be `1`
`weight_decay`	λ constant used in weight decay	Fixed to be `1e-4`
`dropout`	Dropout ratio during training	Fixed to be `0.1`
`alpha1`	The α weight balancing the two-component loss function	`tune.loguniform([0.1, 10])`

Compute Resources. All training sessions of RADT models were done using a single H100 GPU. On average, RADT takes around 10 hours (i.e., 10 GPU hours) to converge on a strong policy on Gymnasium Robotics tasks (depending on the task) in terms of pure *training* time. The *total* amount

of time depends on how often we perform validation/evaluation (e.g., performing 100 interactive episodes with the environment during each evaluation session can take quite a bit of time). The maximum memory utilization during training hit 40 GB. We acknowledge this is a higher computation overhead compared to baselines, which can converge on a strong policy using 1 GPU within 2 hours of pure training time, with the maximum memory utilization during training hitting 20-30 GB. However, given that all training was done with a single GPU and a strong policy can be achieved within a day, we do not think RADT is prohibitively expensive to train. Additionally, RADT has the additional benefit of being able to generalize zero-shot to novel avoid-region configurations dynamically specified at test time while other methods do not, removing the need for compute-heavy retraining in many scenarios where baselines would have to be retrained.

C Experiment Details

C.1 GymnasiumRobotics Environments

Choice of Environments. While there are more complicated environments that are part of the Fetch and Maze suites in Gymnasium robotics, we do not evaluate on those, since the focus of those environments is testing for proficiency in long-range planning, skills learning, and hierarchical learning, which are not the focus of the current iteration of RADT. Additionally, we are focusing only on the domain where the training data is 100% generated from a random policy (Criterion 1.2), and it is difficult for RL approaches across the board to just learn good goal-reaching performance on these more complex environments under this data regime [43], let alone good avoid behavior. With low goal-reaching success rates, our evaluation of avoid behavior will not be very meaningful. As an extreme example, a policy that makes the agent stay in place will achieve an MNC of 0, but its SR will also be 0; this would not be considered good avoid behavior despite the low MNC.

Passable and Impassable Avoid-Regions. In our custom environments, we make the avoid-regions passable (i.e., the agent can pass through these avoid states). This is in contrast to physical boxes that provide a hard constraint on the state space (i.e., the agent cannot cross the boundaries of the avoid-box) as is used in [6]. Hard constraints on the state space present an issue: no training trajectories generated in this environment truly violate any $\mathbf{b}_j$, because no training trajectories actually pass *through* any $\mathbf{b}_j$ (Figure 3b). Thus, all trajectories are somewhat optimal, in the sense that they never demonstrate "unsuccessful" avoid behavior. This setup thus fails to demonstrate Property (2.2) in Section 2. We argue that this is an issue because it couples successful goal-reaching behavior with successful avoiding behavior. If training trajectories cannot go through avoid-boxes, all training trajectories that have succeeded in reaching the goal must have done so by circumventing avoid-boxes. A training trajectory where the agent is adamant on attempting to go through an avoid-box will get stuck at the box boundary and fail to reach the goal. Therefore, with a hard constraint setup, goal-reaching ability is tightly coupled with avoid ability, while we wish to *isolate* these two aspects and see if a model can learn to acquire both goal-reaching and avoid abilities when it is possible to only acquire one but not the other. While this may not be important in the specific context of these Gymnasium Robotics tasks, in general, there may be application scenarios in which we would like to avoid a region of the state space that the agent is technically *able* to pass through, but we would prefer it not. For example, there may be a road in which a self-driving vehicle *can* pass through or *has* previously passed through, but we would rather it not on some particular day due to construction traffic.

Environment Specifications. For the baseline models (AM-Lag, RbSL, WGCSL), the `FetchReachObstacle` environment state space consists of 19 dimensions. The first 10 dimensions correspond to the state space of the original `FetchReach` environment in Gymnasium Robotics [9] and describe information about the location/orientation of the robot arm and the environment in general. The next 9 dimensions describe information about the single obstacle/avoid-region present in the environment, with the same setup as used in [6]. For RADT, the `FetchReachObstacle` environment just has the first 10 dimensions (i.e., the same state space as the original `FetchReach` environment), as RADT does not utilize an avoid-region-augmented state space. For the baseline models, the `MazeObstacle` environment state space consists of $4 + 2 * n_{\text{avoid}}$ dimensions. The first 4 dimensions correspond to the state space of the original `PointMaze` from Gymnasium Robotics and describe the location and velocity of the agent. Then, for each avoid-region in the environment, there are an additional 2 dimensions added to the state space representing the xy-location of the centroid. Thus, the state space dimensionality gets larger the more avoid-regions the environment is

specified to have. For RADT, however, the state space of `MazeObstacle` is just the state space of `PointMaze` and contains 4 dimensions. The avoid-regions in the `MazeObstacle` environment are circles of radius 0.2 around the avoid centroids.

Each environment has a parameter `max_episode_steps`, which dictates the maximum number of timesteps in an episode of interaction with that environment. An episode ends either when the `max_episode_steps` number of timesteps is reached or if the agent successfully reaches the goal. `max_episode_steps` is set to be `50` for `FetchReachObstacle`, `300` for `UMazeObstacle`, and `800` for `AMazeObstacle` and `HMazeObstacle`, which correspond to the original default values of `max_episode_steps` for `FetchReach` and `PointMaze_UMaze` from Gymnasium Robotics.

Length-Normalized Cost Return. We observe in the visualizations of our preliminary experiments that it is possible to "hack" absolute cost return by attempting to reach the goal state in as few timesteps as possible and rushing directly through the avoid-region (Figure 7). Such a policy, which we shall refer to as the "rushed policy," can achieve low absolute cost by minimizing the number of timesteps in the trajectory in total, and thus also minimizing the absolute number of timesteps spent in the avoid-region. However, for reach-avoid applications where safety can be critical, the rushed policy is not preferred to a slower, more cautious policy (which we shall refer to as the "cautious policy") that may take more timesteps to reach the goal but demonstrates a stronger attempt at circumventing the avoid-region. The cautious policy may result in a trajectory that accumulates a similar absolute cost return as the rushed policy trajectory (because it may have skimmed the avoid-region for a similar absolute number of timesteps as the rushed policy spent in the avoid-region), but because the overall cautious policy trajectory makes a better attempt at circumventing the avoid-region, it achieves a much lower length-normalized cost return compared to the rushed policy trajectory. That is, a smaller *proportion* of the timesteps in the cautious policy trajectory violate the avoid-region compared to the rushed policy trajectory, indicating higher quality avoid behavior compared to the rushed policy trajectory. In safety-critical contexts, as long as the agent can reach the desired goal in a reasonable time (i.e., the specified `max_episode_steps`), the *quality* of avoid behavior is more important than the speed of reaching the goal. Figure 7 demonstrates this with two visual examples.

Avoid-Region Relabeling. For training data from the `FetchReachObstacle` environment, hindsight-relabeled avoid centroids are sampled according to the contour-based sampling method described in A.3. For training data from the `MazeObstacle` environments, hindsight-relabeled avoid centroids are sampled using the naive uniform strategy, but only from the parts of the state space that are accessible by the agent; i.e., avoid centroids cannot be sampled inside the walls of the mazes. This is to prevent the model from conflating dynamically specified avoid-regions in the prompt with hard constraints on the state space that are inherent to the environment. Since the avoid-regions in `MazeObstacle` environments are circles of radius 0.2, we choose the circumscribing box of width 0.4 around the avoid centroid for our avoid-box representation in RADT, a conservative choice.

Baselines Cannot Generalize Zero-shot. Here, we clarify in more concrete examples why zero-shot generalization to avoid-box sizes and numbers is not feasible with AM-Lag and RbSL as they are set up in [6].

For the `FetchReachObstacle` tasks, assume at training time the avoid-boxes b in the environment have width w. The augmented state space only includes information about the center of the avoid-box: $\texttt{centroid}(b)$. The *size* of the avoid-box is encoded by the cost function. The cost function that AM-Lag and RbSL are trained on outputs a cost of 1 if the agent's state is within the box of width w surrounding $\texttt{centroid}(b)$, otherwise, it outputs 0. Now, say we increase the avoid-box sizes to $2w$ at evaluation time. The only information the agents get about the avoid-region at evaluation time is in the centroid information $\texttt{centroid}(b)$ present in the state space, which carries no information about the box size. The agents are still trained to stray away from the box of width w surrounding $\texttt{centroid}(b)$, as this is the cost function they are trained to minimize the value of. Thus, to generalize to this new box size, we must define a new cost function that outputs a cost of 1 if the agent's state is within the box of width $2w$ surrounding $\texttt{centroid}(b)$, then *retrain* the agents on this new cost function.

For the `MazeObstacle` tasks, assume at training time that there are 3 avoid-regions b_1, b_2, b_3 in the environment. The augmented state space for AM-Lag and RbSL at training time thus has dimensionality $4 + 2 + 2 + 2 = 10$, with 2 additional dimensions for each of $\texttt{centroid}(b_1), \texttt{centroid}(b_1), \texttt{centroid}(b_1)$ (see Environment Specifications above). Say, at eval-

uation time, we increase the number of avoid-regions to 5: b_1, b_2, b_3, b_4, b_5. Now the state space dimensionality is $4 + 2 + 2 + 2 + 2 + 2 = 14$. We will have to retrain the agents on this new state space.

C.2 Cardiogenesis Environment

Cardiogenesis Boolean Network Details. The boolean network model comprises 15 genes, each represented by a binary variable indicating expression (1) or non-expression (0), yielding a discrete state space of size 2^{15}. The model is depicted by Figure 8a and described mathematically in [49, 20]. Actions correspond to genetic perturbations that flip the expression value of a single gene. After perturbation, the Boolean network to a new gene expression state based on its internal logic. Updates are done asynchronously, as is done in [49]: a random gene node is chosen, and its value is set (either changed or remained in place) such that it satisfies all boolean rules . A *stable attractor state* is defined as a state in which all boolean rules in the boolean network are satisfied, such that the boolean network is self-consistent and further asynchronous updates will not change the node values (assuming absence of an external perturbation).

In our `Cardiogenesis` environment simulations, a state-action-state transition sequence is obtained as follows: 1) we start with the boolean network representing the current gene expression state $\mathbf{s_t}$, 2) an action is chosen (i.e., a chosen gene node is perturbed and has its value flipped from `0` to `1` or vice versa) to create a post-perturbation *transient state*, 3) `k` asynchronous updates to the boolean network are performed to the network, and 4) the resulting state after `k` asynchronous updates is defined to be the next state in the trajectory, $\mathbf{s_{t+1}}$. This is depicted in Figure 8b. Since the boolean network updates are asynchronous, there is some stochasticity involved in the transitions due to the random selection of genes to be updated at each asynchronous update step. The value of `k` affects how stochastic our transitions are; the higher the value of `k`, the more likely the boolean network model will hit a *stable attractor state* in the process, and the less noisy the transitions of our `Cardiogenesis` environment will be. We choose `k=10`. Note that this value of `k` does not guarantee that all states $\mathbf{s_t}$ will be stable attractor states; this is intentional, since in reality, a cell can be perturbed while it is unstable and still in the process of reaching a stable attractor state.

Avoid-Region Relabeling. At training time, the offline dataset of 60,000 random policy steps is split into trajectories of length 30. Hindsight goal relabeling is done as usual. During hindsight avoid-region relabeling, we sample avoid-regions from the top 8 most represented states (attractors and non-attractors) in the offline dataset. This choice results in approximately half of all training trajectories being successful demonstrations of avoid behavior on the first pass of hindsight avoid-region relabeling, removing the need to do a second pass. Since we are working with a discrete state space, avoid states that are discrete states rather than regions of a continuous state space, and the corresponding avoid-region token representation $\mathbf{b} \in \mathbb{R}^{15}$ is simply the avoid state itself.

Start State/Goal State Sampling. At evaluation time, start states are randomly chosen from stable attractor states. The goal state is fixed as the FHF state (`100010010101100`) for our experiment.

Example Trajectories from Experiments. Here, we show some example trajectories depicting the three scenarios described in Section 5. Note that, unless otherwise specified, we depict trajectories where repeated states are collapsed, which we call the "collapsed trajectory." For example, the trajectory `(a, a, b, c, d, d, d)` would be collapsed into the collapsed trajectory `(a, b, c, d)`. It is important to note that it is possible for an action to lead to staying in the same state due to the dynamics of the Boolean network model. These trajectory examples are also organized in Table 5.

As an example of Scenario 1, when the starting state is `000000000000000`, RADT's policy, when given no avoid prompt, always leads to a collapsed trajectory:

`(000000000000000, 100000000001100, FHF)`

However, upon adding `100000000001100` as an avoid prompt, the resulting policy always leads to a collapsed trajectory:

`(000000000000000, 000011010100000, 000010010100000, FHF)`

which is a longer, alternative path that deterministically circumvents `100000000001100`.

Table 5: Example trajectories from the `Cardiogenesis` environment

Initial state	Most visited intermediate	Examples
		- avoid prompt:
		(000000000000000, 100000000001100, FHF)
		+ avoid prompt:
000000000000000	100000000001100	(000000000000000, 000011010100000, 000010010100000, FHF)
		- avoid prompt:
		(010111111010011, 001111011011111, FHF)
		(10111111010011, 010111111010011, 001111011011111,
		...000111010011111, 001111011011111, 100000000001100, FHF)
		(010111111010011, 001111011011111, 000111010011111, 001111011011111, FHF)
		+ avoid prompt:
		(010111111010011, 000000000000011, 010010010100000,
		...010111111010011, 000010010100000, FHF)
010111111010011	001111011011111	(010111111010011, 000010010100000, 010111111010011, 000010010000000, FHF)
		- avoid prompt:
		(000010010100000, 000010010100000, 000010010100000,
		...000010010100000, 000010010100000, FHF)
		+ avoid prompt:
000010010100000	000010010100000	(000010010100000, 000010010100000, 000010010100000, FHF)

In Scenario 2, the stochastic environment plays a much more prominent role. For example, when the starting state is 010111111010011, and RADT is given no avoid prompt, the resulting policy leads to a variety of trajectories. A few examples are:

(010111111010011, 001111011011111, FHF)

(10111111010011, 010111111010011, 001111011011111, 000111010011111, 001111011011111, 100000000001100, FHF)

(010111111010011, 001111011011111, 000111010011111, 001111011011111, FHF)

Upon adding 001111011011111 as an avoid prompt, we still get some similar trajectories involving 001111011011111 as an intermediate state, but we get a higher frequency of trajectories that manage to avoid it, such as:

(010111111010011, 000000000000011, 010010010100000, 010111111010011, 000010010100000, FHF)

(010111111010011, 000010010100000, 010111111010011, 000010010000000, FHF)

The takeaway here is the diversity of trajectories that are possible given the same policy due to the stochastic nature of the environment.

As an example of Scenario 3, when the starting state is 000010010100000, all trajectories have collapsed form:

(000010010100000, FHF)

when there is no avoid prompt, meaning the cell can directly reach the goal state from the start state without passing through a third intermediate state. However, the uncollapsed form of these trajectories has length 5 and is:

(000010010100000, 000010010100000, 000010010100000, 000010010100000, 000010010100000, FHF)

indicating that RADT makes multiple attempts to get out of 000010010100000 to land on FHF and only manages to get out on the 5th attempt, spending a total of 5 timesteps stuck in the initial state 000010010100000. However, when we add the initial state 000010010100000 as an avoid prompt (which is unavoidable since it is the starting state), the uncollapsed form of all resulting trajectories has length 3 instead and is:

(000010010100000, 000010010100000, 000010010100000, FHF)

Here, RADT cannot avoid the avoid state 000010010100000 entirely, but it is encouraged and finds a way to spend less time at the 000010010100000 state, getting out of the state more quickly (40% fewer attempts).

Baselines. Additionally, while our primary baselines (AM-Lag and RbSL) are DDPG-based actor-critic models that are designed to work specifically with continuous action/state spaces, we create naively discretized versions of the baselines (i.e., adding a softmax layer to the actor model for

outputting discrete actions) and run them on 'Cardiogenesis'. We recognize that these discretized versions are not representative of the original methodologies and that more sophisticated method of translating them to discretized spaces exist, though they are not trivial and thus highlight the lack of adapability of these approaches to different state/action space structures. As we suspected, the results are quite poor as shown in Table 16. Whilst RADT was able to achieve 100% goal-reaching SR at all times (with or without avoid prompts), AM-Lag and RbSL SRs are quite low, making it difficult to properly compare avoid behavior with RADT. Even when RbSL manages to have a 100% SR in Scenarios 1 and 3, as soon as we add the avoid prompt of the most visited intermediate, SR drops to 0%. MNC drops to 0% as well, but this is not meaningful, as the agent could just be sitting still or wandering aimlessly in the state space. It is difficult to tell whether the poor performance is attributed to the baselines' inherent capabilities, or to the poor translation of a continuous model into a discrete model. Regardkess, this does further emphasize RADT 's advantage in being much more adaptable to different state/action space structures compared to baselines.

C.3 Other Notes

Lidar-based Observation Spaces and Goal-Conditioning. Regarding the discussion about OSRL approaches in Section 3, it is worth noting that the standard "reach avoid" benchmarking task for OSRL algorithms, the `Safe Navigation` environment from Safety Gymnasium [47, 17], may seem like it is testing the goal-conditioning and avoid-region-conditioning abilities of OSRL models, but it does not require true goal-conditioning. This is because it utilizes a relative-perspective, lidar-based observation space $\mathcal{S}_{\text{lidar}}$ that allows the agent to only have to learn to reach *one* state in $\mathcal{S}_{\text{lidar}}$ in order to "generalize" to any target location in physical space: the $\mathbf{0}$ vector (indicating that the agent is 0 distance away from the goal). Therefore, models do not need to generalize to any arbitrary goal in the observation space $\mathcal{S}_{\text{lidar}}$, and thus OSRL models do not need to be truly goal-conditioned in $\mathcal{S}_{\text{lidar}}$ (most are not). The same logic can be applied to argue that the tasks do not check for true avoid-region-conditioning in the observation space. While relative-perspective observation spaces are advantageous in this regard, they cannot be conceived or used in all environments, e.g., cell reprogramming.

D Additional Experiments

D.1 Ablation Experiments

All ablation studies are performed in the `FetchReachObstacle-BoxWidth0.16` environment, and the AS/GS tokens ablation study was additionally done in the `UMazeObstacle` environment with 3, 5, and 7 avoid states.

Ablation of the SI token. As shown in Table 6 in the Appendix I, RADT with the SI token (+SI) does drastically better in terms of both MNC and SR compared to RADT with the SI token removed and trained on the full dataset of both successful and unsuccessful avoid examples (-SI, full data). This is as expected, since without the SI token, the model has no signal to differentiate what is considered successful and unsuccessful avoid behavior in the training examples. Additionally, +SI also drastically outperforms RADT with the SI token removed and trained on a dataset *only* consisting of successful avoid examples (-SI, successes only). This clearly demonstrates the importance of including both failure examples and the SI token to teach the model what is *undesirable*.

Ablation of the AS/GS tokens. As shown in Table 7 in Appendix I, RADT with the AS and GS tokens (+AS,GS) performs notably better in terms of both MNC and SR compared to RADT without the AS and GS tokens (-AS,GS) in `FetchReachObstacle`. In the `UMazeObstacle` environments, we observe that MNC is worse in the -AS,GS experiments than in +AS,GS setup (approaching MNC values closer to those seen with the WGCSL baseline), but there is no significant difference in SR. The performance drops in terms of MNC are not as drastic compared to those seen with the -SI ablation experiments, as -AS,GS model may potentially be able to learn that only the last token in the prompt is considered a goal state regardless of how many tokens precede it. These results show that the AS and GS tokens are not as critical as the SI token, but do provide noticeable benefits.

Ablation of avoid-region tokens and size component. We also run experiments in the `FetchReachObstacle` environment where we 1) ablate the avoid-region (AR) tokens while leaving the indicator tokens and 2) ablate only the *size component* of the AR token vectors (i.e., only including centroid location information, no box or radius representation) to isolate the importance

of including size/extent information in our avoid-region token representation. As shown in Table 8 in Appendix I, both ablations drastically worsen MNC across all box sizes.

Ablation of attention boosting. We find that adding a bias `adelta` (a hyperparameter) to the attention logits to all prompt tokens, using the strategy described in [48], noticeably improves the instruction-following ability of RADT to the prompt (Figure 6 in Appendix I).

Ablation of avoid-region relabeling. As shown in Table 13 in Appendix I, the complete ablation of avoid-region relabeling (i.e., using the avoid-regions that were originally generated by the environment during the collection of training data) significantly impairs RADT's MNC performance across both ID and OOD box sizes, demonstrating the critical nature of avoid-region relabeling in `FetchReachObstacle`. Note that goal relabeling and the avoidance success indication (SI token) are left intact in these experiments; therefore, the goal-reaching SR is not significantly impacted.

D.2 Extremities of OOD Generalization

To test the limits of RADT's zero-shot OOD generalization, we train RADT on `UMazeObstacle` trajectories with maximum 3 avoid-regions and evaluated on `UMazeObstacle` environments with 10, 15, and 20 avoid-regions. We also evaluate the strongest baseline, AM-Lag, on these environments, but like with the main experiments, AM-Lag is retrained on each new environment (i.e., ID, not zero-shot) which gives it an advantage over RADT, and thus it represents the "ideal" performance. The results included the results in Table 9.

At 10 avoid-regions, RADT 's zero-shot OOD performance is still on par with the retrained AM-Lag model in terms of MNC and better than the retrained AM-Lag model in terms of SR. At 15 and 20 avoid-regions, however, we see that RADT 's zero-shot avoid behavior starts to lag behind the retrained AM-Lag's according to MNC, indicating that RADT 's OOD performance does indeed have a limit (MNC suffers when the number of avoid-regions is over 300% of the maximum number of avoid-regions seen during training). However, interestingly, the zero-shot RADT policy results in a SR that is still higher than the retrained AM-Lag policies. This is a strong result because retraining AM-Lag ensures that scenarios with a larger number of avoid-regions become in-distribution settings for AM-Lag, whereas they remain out-of-distribution for zero-shot RADT.

Effects on attention mechanism. Additionally, we qualitatively analyze the effects of increasing prompt length (due to increasing the number of avoid-regions) on the quality of the attention mechanism. Specifically, we check to make sure that long prompt lengths do not result in critical attention defects such as consistent attention sinks or uniform/zero attention.

For each episode during evaluation, we calculate the attention weights, averaged across all attention heads, that the model gives to the first 30 tokens of the input sequence (including both the prompt and the early state/action tokens in the trajectory history) when generating the *last* action token of the trajectory. We plot these values for 1, 5, 15, and 20 avoid-regions (Fig. 11). We choose to analyze the attention mechanism at the last action token as this is the most challenging token in terms of long-range dependence on the prompt. We notice no major attention sinks or debilitating attention uniformity across the prompt tokens that result from long prompt lengths, as there is still healthy diversity in attention distributions across different trajectories even when there is a very large, OOD number of avoid-region tokens in the prompt (e.g. 15-20).

However, we do note a few interesting observations. First, RADT pays very little to no attention to the state/action tokens in the early portion of the trajectory history (occurring right after the prompt tokens) when generating the last action in the trajectory. This is likely because the model has learned that it is sufficient to make action predictions with a Markovian-like policy, taking into account only the more recent states while ignoring historical states from earlier parts of the trajectory history. Second, while there are no major attention sinks and uniformity observed, we do notice a drop in attention to individual avoid-region tokens on *average* with very OOD prompt lengths (e.g. 15 and 20 avoid-regions). Intuitively, this is to be expected. It is more challenging for a model to keep track of a high number of different avoid-regions, so it must be selective in distributing its attention only to the most relevant ones at any particular action; it cannot afford to constantly pay attention to every single avoid-region like it can when there are only a couple. Additionally, this would be consistent with our discussion in Appendix G.4 "More complex avoid-region shapes." As avoid-regions increase in the `MazeObstacle` environments, avoid-regions increasingly start to overlap with one another to compose effectively larger avoid-regions, so it would make sense that the significance of individual avoid-regions would decrease given the large amounts of overlap. While

we acknowledge the decrease in average attention to individual avoid-region tokens with longer prompts can be a limitation, it is much more likely that the difficulty of the task itself would limit the number of avoid-regions before issues related to attention across long contexts arise. For example, 20 avoid regions in the maze environment already makes the environment way too crowded. Third, RADT always puts a lot of attention weight on the *goal token* at the end of the prompt, regardless of how many avoid-region tokens precede it or how long the prompt is as a whole. This is not surprising, as the goal token is likely the most relevant token at the last action of the trajectory, and there is ever only one goal token. However, it does show that RADT is robust to increasing prompt length, as it is still able to identify the critical nature of the goal token despite the many more potentially "distracting" tokens in longer prompts.

D.3 Effects of Expert-policy Data

To examine the performance ceiling induced by only using random-policy data, we run an experiment in the `FetchReachObstacle` environment where we train RADT on a mix of random-policy and expert-policy data provided by the RbSL study [6] at 4 levels (0%, 10%, 20%, and 30% expert-policy data). The MNC and goal-reaching SR results are included in Table 10.

Looking at the MNC values, we see that introducing expert-policy data seems to have little effect on avoid performance for ID box sizes (box width 0.16) and slightly OOD box sizes (box width 0.18). This suggests that, for ID box sizes, the performance ceiling of RADT 's avoidance behavior when using only hindsight-relabeled, random policy training data is close to what RADT can ideally achieve by learning from expert demonstrations (i.e., using only hindsight-relabeled, random-policy data provides no clear limitation).

However, when we get to larger, more OOD box sizes (box width 0.20, 0.22, and 0.24), we see that introducing expert-policy data increasingly starts to make a difference and improve avoid performance. This suggests that using only hindsight-relabeled, random policy training data can limit performance ceiling on RADT for *larger, OOD* box sizes (i.e., RADT can generalize better in a zero-shot manner to larger, more OOD box sizes if provided with expert data). A possible explanation for this is that the expert-policy demonstrations provided by the RbSL study are more conservative trajectories (i.e., widely avoiding the box with a much larger spatial buffer) than hindsight-relabeled, random-policy trajectories. When provided with these demonstrations involving large, wide, conservative arcs to avoid the box, RADT can better generalize to larger box sizes even when the box sizes themselves are technically OOD.

Interestingly, we also observe that the goal-reaching SR values are *adversely* impacted by introducing expert-policy data across all box sizes, and more prominently with larger, more OOD box sizes (box width 0.2, 0.22, 0.24). In fact, with only random-policy data, RADT maintains a high SR across all box sizes, but when expert-policy data is introduced, SR deteriorates with increasing box size. A possible explanation for this is that the cost of the improved avoid performance for larger box sizes provided by the inclusion of expert-policy data is a deterioration in goal-reaching ability. This is expected as larger box sizes require longer trajectories to successfully circumvent, so given that the allotted time to reach the goal is fixed across all experiments, it is more likely that the agent does not successfully reach the goal in time when it is successfully circumventing a very large box.

We also provide baseline results for this expert-mixture experiment. The results are shown in the Table 10; as usual, it is important to note that the baselines are retrained for every box size while RADT was applied zero-shot. As with the main results in Table!1, RADT beats both baselines in MNC across all settings except for the box width 0.16 settings, where RADT loses to AM-Lag on MNC but has superior SR. In general, we notice the same trend as we saw with RADT, where increasing the proportion of expert data generally improves MNC while sacrificing SR, implying that the bias that expert demonstration brings is to be more conservative. However, the drop in SR across expert % for box sizes 0.2-0.24 is much less drastic for the baselines than for RADT, which we attribute to RADT being applied zero-shot and the baselines being re-trained on every setting. It thus seems that the data bias introduced by expert data is exacerbated the more OOD the zero-shot setting gets.

D.4 Effects of Changing Avoid-Region Token Representation

To examine how flexible RADT is with regards to the vector representation used for avoid-region tokens $\mathbf{b}_j \in \mathbb{R}^{d_a} : j \in \{1, 2, ..., n_{\text{avoid}}\}$, we train and evaluate RADT on the `UMazeObstacle` environment using the alternative representation for spherical avoid-regions, as described in Appendix A.2. This representation should allow the model to learn a tighter bound around the avoid-regions, compared to the more conservative circumscribing box representation described in Appendix C.1.

As the results in Table 12 show, changing the avoid-region token representation from the box representation to the spherical representation does not impact RADT's to learn strong reach-avoid policies, with performance that is comparable to the original `UMazeObstacle` experiments. One difference is that using the performance using the spherical representation seems to lead to *slightly* worse MNC and a higher SR; this is consistent with intuition, as the original circumscribing box representation would provide a more conservative boundary of the avoid-regions in `UMazeObstacle` and thus result in more conservative trajectories that have better MNC at the cost of goal-reaching SR.

More complex avoid-region shapes. In addition to being flexible to different avoid-region token representations, RADT can be prompted to avoid more complex avoid-regions by representing them as a composition of multiple smaller avoid-regions. This is implicitly demonstrated in the `MazeObstacle` experiments with a high number of avoid-regions (Section 5 Results 2, Section 5 Results 4 "Extremities of OOD generalization", Appendix D.2). In these settings, the high number of randomly generated avoid-regions will result in groups of partially overlapping avoid-regions that compose together to create effectively larger avoid-regions of complex shapes (Fig. 10).

D.5 Simultaneous Varying of Multiple Avoid-Region Axes

To demonstrate RADT's ability to respond to variation in multiple axes of avoid-region specification simultaneously, we implement a navigation task called `MotionPlanning`, where the agent must reach a goal in an open rectangular environment while avoiding a random number (ranging from 2 to 4) of smaller rectangular avoid-regions, each generated with random location, aspect ratio (ranging from 1:1 to 1:2.5), and size (with side lengths ranging from 0.24 to 0.6). Unlike previous tasks, this environment simultaneously involves variation in location, shape, size, and number of avoid-regions. RADT and baselines were trained in-distribution on this task, and the results are in Table 14. As shown, RADT achieves the best MNC on this task and is on par with AM-Lag for SR, with RbSL lagging far behind.

D.6 Image-Based Observation Spaces and Avoid-Region Representations

To explore RADT's potential in handling 1) high-dimensional observation spaces and 2) unstructured, flexible avoid-region token representations that can capture abstract avoidance desires without knowledge of a pre-defined vectorization, we conduct a preliminary proof-of-concept demonstration of RADT applied to an image-based version of `UMazeObstcale` that we call `UMazeObstacleImage`. Here, we replace the observation space for `UMazeObstacle` with an image-based observation space. To obtain state $\mathbf{s}_t$, we take an 224x224 image of the maze with the agent's current position visualized as a green sphere, then pass the image through Torchvision's pretrained ResNet-18 model [19, 39]. We then take the representation outputted by the fully-connected layer at the end of the ResNet-18 architecture as our state representation, giving us a 1000-dimensional state vector $\mathbf{s}_t \in \mathbb{R}^{1000}$ (Fig. 12). Similarly, we acquire avoid-region tokens and goal tokens via the same process, except we replace the green sphere representing the agent's position in the original image with blue objects representing the avoid-regions and red spheres representing goals, respectively. Therefore, the avoid-region tokens $\mathbf{b}_j \in \mathbb{R}^{1000}$ and the goal token $\mathbf{g} \in \mathbb{R}^{1000}$ have the same dimensionality as the state vectors.

We then use the same hindsight-relabeled training trajectories from our `UMazeObstacle-NumAvoid1` dataset for training, but replace all the state, avoid-region, and goal tokens with these new representations. We benchmark against AM-Lag, RbSL, and WGCSL as baselines using the same evaluation process described for the `UMazeObstacle` environments. As shown by the results in Table 15, RADT achieves superior MNC and comparable or superior goal-reaching SR to all baselines. The SR values here are lower than the SR results from the

original `UMazeObstacle-NumAvoid1` task, but this is to be expected given the much more difficult nature of a high-dimensional state space. The SR values are also within the range of typical SR values that SOTA GCRL approaches achieve on image-based, goal-reaching maze tasks (that do not include an avoidance objective), according to the results from OGBench (which do not involve an avoidance component) [43]. While there is room for additional improvement through further optimization of hyperparameters, this preliminary study demonstrates RADT's strong potential in being adapted to both high-dimensional tasks and flexible avoid-region representations that do not require precise knowledge of a strict, vectorized structure to specify avoid-regions.

E Extended Related Work

Below are some additional discussions regarding the categories of related work described in Section 3 and their limitations with regard to satisfying the ideal properties described in Section 2

Offline Goal-Conditioned RL. While reward-based OGCRL can technically satisfy Property (4) by designing a reward function that takes into account both goal information and avoid-region information (e.g. positive reward for approaching the goal, negative reward for approaching an avoid-region), they *effectively* fails to satisfy Property (4) in practice. While it is theoretically possible to express information about both desirable and undesirable states in a multi-component reward function, such functions are practically difficult to design such that both these sometimes conflicting desires are balanced properly [24, 25, 12]. Also, the reason why reward-free OGCRL algorithms strictly do not satisfy Property (4) is because, without a reward, there is no way to specify any additional desires outside of the goal $\mathbf{g}$, such as avoid-regions. Thus, reward-free OGCRL is usually a good fit in situations where the path to the goal does not matter or the demonstrations used for training are expert/optimal demonstrations that take an ideal path to the goal.

An argument can be made that reward design for reward-based methods are trivial in simpler environments and goal-reaching tasks where we can just define a sparse reward function for reaching the goal. Regarding this, reward-free GCRL has one other major advantage: sample-efficiency. The referenced work [5] formally explores why reward-free, self-supervised methods (e.g., contrastive learning methods, DT-based methods like ours, etc.) are more sample-efficient than TD-based methods (e.g., Q-learning variants) with a sparse reward function. The intuition, as explained in [5], is that TD-based methods in sparse reward settings do not get any learning signal until a reward is observed, and when a reward is observed, credit assignment among the various state transitions in the trajectories is difficult to disentangle without large amounts of data. Self-supervised approaches, on the other hand, can learn from every individual state transition (by learning how to reach every successor state in every state transition). That is, they work by implicitly learning state succession dynamics with every state transition data point; therefore, self-supervised approaches can be considered in between model-based (explicitly learning environment dynamics) and model-free RL, and are therefore more sample-efficient for the same reasons why model-based RL is often more sample-efficient.

Offline Safe RL. OSRL approaches circumvent the challenge of having to design multi-component reward functions by using a separate cost function to capture information about avoid behavior and solving a constrained Markov decision process [3], isolating the handling of goal-reaching and avoiding desires. We claim in Section 3 that OSRL algorithms are not truly goal-conditioned in general; however, this may seem surprising since a very common benchmark for OSRL algorithms is a reach-avoid environment called `Safe Navigation` from Safety Gymnasium [47]. However, we explain in Appendix C.3 why the tasks in this environment are not true multi-goal RL tasks.

***Online* Goal-Conditioned Safe RL.** While the RbSL paper [6] (and the associated algorithms RbSL and AM-Lag) is one of the only pieces of literature we have found that explicitly attempts to create an approach that is *offline*, goal-conditioned, and avoids region-conditioned, we acknowledge that there exist other goal-conditioned, avoid-region-conditioned algorithms designed for reach-avoid tasks that are *online* algorithms [11, 50, 21]. Online algorithms may work in domains like robotics, where we can reasonably create safe testing environments or have good simulators, but we would like to create an approach for domains in which online training is not feasible.

Decision Transformers. We build our model off of MGPO, which is truly goal-conditioned and thus satisfies Property (3). During the offline pretraining phase, MGPO is purely data-driven and does not use a reward function, fully relying on hindsight goal relabeling. However, MGPO does use an online prompt optimization phase in addition to offline pretraining (failing to satisfy Properties (1)

and (2)), utilizes reward functions during online finetuning (ultimately failing to satisfy Properties (5)), and does not explicitly take into account avoid-region information, unless it is baked into the reward function design (effectively failing to satisfy Properties (4)). Our work builds upon MGPO such that it is more ideal for reach-avoid tasks, aiming to satisfy these remaining properties.

JEPA-based models. There have been recent JEPA (Joint-Embedding Predictive Architecture)-based planning methods that can also tackle sequential decision making problems in a reward-free, suboptimality-tolerant, and offline manner [51, 69]. However, the primary difference between such planning methods (which are more aligned with a model-based RL paradigm) and model-free RL methods like RADT is test-time/deployment-time compute. Planning methods require exploring, searching through, and rolling out many possible plans at test time, and avoidance checking would thus happen at test time (i.e., used to select appropriate plans in the search), requiring much higher test-time compute. Model-free methods like RADT, on the other hand, front-loads avoidance checking to training (i.e., to calculate the values for the training examples); at deployment time, a single rollout is made (no searching) and no avoidance checks need to be performed. Additionally, while it is possible to add an avoid-region violation penalty to the planning cost to make JEPA-based planning models have avoidance capabilities, there are a couple of fundamental challenges here that RADT does not face. First, since the primary innovation of JEPA-based models is that planning is done in the latent space, this avoidance check would have to be performed at test time in the latent space. Checking violation of the exact avoid-region boundaries in the state space would be difficult to do in the latent space, which does not necessarily preserve state space geometry, and decoding back into the state space to perform this check would be computationally expensive and undermine the point of JEPA-based approaches. Second, introducing an avoid-violation penalty term in the planning cost function would once again re-introduce the problem of having to manually tune the balance between the possibly conflicting objectives of reach and avoid (Property (5)). Third, models like Dino-WM are designed to work strictly in a continuous state space with no clear way to cleanly translate such models to a discrete space like Cardiogenesis (Section 5); such adaptability to different state space types is a major advantage of RADT.

Pareto vs. Constrained Optimization Evaluation Framework. Many of the safe RL methods discussed above frame the problem as a Constrained Markov Decision Process (CMDP), which is inherently a constrained optimization problem framing rather than a Pareto optimization framing like ours. In a constrained optimization framing, one objective (usually the avoidance objective in reach-avoid problems) is treated as a constraint that must be satisfied in order for the other objective (the reach objective) to even be considered when determining success; for instance: $\max_\theta \ J_{\text{reach}}(\theta) \quad \text{s.t.} \quad \mathbb{E}_{\tau\sim\pi_\theta(\cdot|g,B)}\big[\, M(\tau, B)\,\big] \leq k$) 1) This implies an ordering/hierarchy of importance and priority between the two objectives. That is, in this formulation, there is always a way to determine which is the best method: first compare constraint satisfaction, if constraint is satisfied, then compare main objective. A Pareto-style evaluation framework is more general and precludes imposing such a priority hierarchy in evaluation and comparison. 2) Under a strict constraint optimization framework, one of the objectives (the constraint) can only be compared on binary granularity on the episode level (either the constraint, like avoiding a region, is satisfied in the episode or it is not), losing nuanced, granular comparison between methods on that objective. This is fitting for scenarios where avoidance is a *strict*, hard constraint. However, in many cases, avoidance is not a strict/hard constraint (e.g. passing through a construction zone in driving navigation is allowed and not destructive, though still undesirable) and sometimes even unavoidable (e.g. Cardiogenesis scenario 3, Section 5). In such cases, when avoid-regions can be or must be violated, how *much* they are violated matters (e.g., how much of your route you spend in a construction zone affects how fast you reach your destination, how much time your cell spends in an undesirable gene expression state affects risk of tumorigenesis), which also relates to our motivation of using MNC as our evaluation metric of avoidance behavior (Appendix C.1).

However, we recognize that there exist reach-avoid applications that do involve strict avoidance constraints and do fit under the priority hierarchy implied by a constrained optimization framing; in these cases, immediately terminating the episode upon violation and considering the run a failure can be a more informative evaluation framework. Therefore, for additional insight, we also run the evaluations for the `UMazeObstacle` and `AMazeObstacle` environments where trajectories are immediately terminated upon violating an avoid-region and considered a failure, and report the failure rate (FR). The results are presented in Table 17. Note that, as usual, results for 3 or more avoid regions are zero-shot for]name but retrained for baselines. As shown, for all `AMazeObstacle` settings and ID/slighlty OOD Umaze settings, RADT maintains comparable or superior FR to the

best baseline (AM-Lag), although AM-Lag starts to beat out RADT on `UMazeObstacle` when the number of avoid-regions becomes more OOD for RADT (≥ 5) but is still ID for AM-Lag. Overall, trends/differences in FR are less drastic between the models than in MNC, so when FR values are similar (and especially when they're high, e.g. > 0.5), MNC on unterminated trajectories becomes more informative as it showcases the difference in the *quality* of the trajectory in scenarios where violation is not fatal/is inevitable.

F Broader Impacts

Foundational methods for learning reach-avoid policies, like RADT, have strong potential in improving the performance of technologies in many application domains, such as general robotics, autonomous vehicles, and bioengineering. While we believe the advancement of these downstream domains can greatly benefit society, we do acknowledge that such technologies can also be used maliciously. RADT and any derivatives should never be used to create autonomous agents with the harmful goals (e.g., self-navigating vehicles that maliciously target individuals) or to induce harmful biological states (e.g., programming cells to pathological cell states). RADT is designed with the intent of enabling the development of *safer* sequential decision making agents that can improve the convenience and health of individuals in our society.

G Limitations

The flexibility and zero-shot capabilities of RADT are balanced out by the fact that it takes more computational resources to train upfront compared to the baselines models, having many more parameters as it is based on a GPT-2 architecture (see "Compute Resources" section under Appendix B.2). However, given that RADT can be trained to achieve a strong policy on 1 GPU in less than day in most cases, and the fact that its zero-shot generalizability reduces the amount of task-specific retraining required, this is not an unreasonable tradeoff.

Additionally, as explored in our stress-testing experiments, there does seem to be an empirical limit to RADT's zero-shot generalization to OOD avoid-region specifications (Section 5 Results 4, Appendix D.2). While the extent of the OOD generalization is impressive, there may be scenarios in which a further extent of generalization is desirable.

We also acknowledge that there may be reach-avoid application domains that do not strictly require all of the Desirable Properties presented in Section 2. For example, in many real-world contexts, we do have a lot of high-quality expert demonstrations to use for training (e.g., for autonomous driving). Therefore, the set of Desirable Properties we propose is more of an idealistic upper bound, such that a reach-avoid learning approach that satisfies all of these properties will very likely be appropriate for *any* reach-avoid problem.

Finally, the scope of our work is to establish the prospect and viability of hindsight avoid-region relabeling (HAR) as a paradigm for reach-avoid learning, recognizing that there are multitudinous opportunities for the development of customized, improved HAR sampling strategies optimized for specific applications and state spaces. For example, while we demonstrated promising potential for HAR and RADT to be applied to high-dimensional, pixel-based observation spaces in a proof-of-concept experiment (Section D.6), we acknowledge that more in-depth, dedicated exploration of the nuanced challenges at this front (e.g., handling first-person pixel-based observation spaces) would be a meaningful complementary line of future research, which could potentially involve non-trivial HAR design optimizations specific to those challenges.

H Code

The code for this project can be accessed at the following GitHub repository: https://github.com/mims-harvard/RADT. Our repository is built on top of the repositories for MGPO [66] and RbSL [6].

I Additional Tables and Figures

Table 6: Results for the success indicator token ablation study for `FetchReachObstacle-BoxWidth0.16` (avg. over 3 seeds).

	+SI	-SI, full data	-SI, successes only
MNC	$0.049_{\pm 0.016}$	$0.629_{\pm 0.018}$	$0.614_{\pm 0.027}$
SR	$0.989_{\pm 0.01}$	$0.921_{\pm 0.016}$	$0.917_{\pm 0.024}$

Table 7: Results for the goal start/avoid start tokens ablation study (avg. over 3 seeds).

Environment	+AS,GS		-AS,GS	
	MNC	SR	MNC	SR
FetchReachObstacle-BoxWidth0.16	$0.049_{\pm 0.016}$	$0.989_{\pm 0.01}$	$0.098_{\pm 0.014}$	$0.892_{\pm 0.055}$
UMazeObstacle-NumAvoid3	$0.045_{\pm 0.009}$	$0.868_{\pm 0.028}$	$0.060_{\pm 0.005}$	$0.882_{\pm 0.033}$
UMazeObstacle-NumAvoid5	$0.087_{\pm 0.005}$	$0.853_{\pm 0.015}$	$0.102_{\pm 0.004}$	$0.865_{\pm 0.043}$
UMazeObstacle-NumAvoid7	$0.127_{\pm 0.01}$	$0.832_{\pm 0.038}$	$0.133_{\pm 0.008}$	$0.878_{\pm 0.003}$

Table 8: Results for the avoid-region token and size component ablation study in `FetchReachObstacle`.

Box width	+AR token		-whole AR tokens		-AR size component	
	MNC	SR	MNC	SR	MNC	SR
0.16	0.049	0.989	0.619	0.9	0.464	0.98
0.18	0.107	0.978	0.747	1.0	0.513	0.92
0.2	0.164	0.983	0.853	0.99	0.655	0.98
0.22	0.247	0.989	0.91	0.99	0.72	0.98
0.24	0.409	1.0	0.943	0.98	0.83	0.98

Table 9: Results for evaluating RADT on extreme avoid-region numbers in `UMazeObstacle` (avg. over 3 seeds).

# of avoid-regions	RADT$^{\#}$		AM-Lag	
	MNC (1e-2)	SR	MNC (1e-2)	SR
10	$\mathbf{16.267}_{\pm 1.002}$	$\mathbf{0.923}_{\pm 0.029}$	$\mathbf{16.2}_{\pm 0.5}$	$0.82_{\pm 0.036}$
15	$23.667_{\pm 0.513}$	$\mathbf{0.92}_{\pm 0.005}$	$\mathbf{21.6}_{\pm 0.656}$	$0.75_{\pm 0.015}$
20	$29.333_{\pm 1.762}$	$\mathbf{0.912}_{\pm 0.01}$	$\mathbf{24.397}_{\pm 1.114}$	$0.79_{\pm 0.026}$

$^{\#}$ = Model capable of zero-shot generalization. Results highlighted in blue are zero-shot results.

Table 10: RADT results for different levels of expert-random data mixture, zero-shot on all box widths over 0.16 (avg. over 3 seeds).

Expert %	0		10		20		30	
Box Width	MNC	SR	MNC	SR	MNC	SR	MNC	SR
0.16	$0.049_{\pm 0.016}$	$0.989_{\pm 0.010}$	$0.053_{\pm 0.004}$	$0.972_{\pm 0.025}$	$0.055_{\pm 0.012}$	$0.967_{\pm 0.029}$	$0.053_{\pm 0.017}$	$0.95_{\pm 0.017}$
0.18	$0.107_{\pm 0.018}$	$0.978_{\pm 0.009}$	$0.096_{\pm 0.003}$	$0.921_{\pm 0.008}$	$0.105_{\pm 0.024}$	$0.959_{\pm 0.009}$	$0.103_{\pm 0.019}$	$0.95_{\pm 0.033}$
0.20	$0.164_{\pm 0.027}$	$0.983_{\pm 0.017}$	$0.146_{\pm 0.017}$	$0.922_{\pm 0.011}$	$0.137_{\pm 0.021}$	$0.928_{\pm 0.054}$	$0.142_{\pm 0.031}$	$0.888_{\pm 0.025}$
0.22	$0.247_{\pm 0.018}$	$0.989_{\pm 0.019}$	$0.203_{\pm 0.032}$	$0.944_{\pm 0.042}$	$0.191_{\pm 0.044}$	$0.894_{\pm 0.034}$	$0.185_{\pm 0.138}$	$0.856_{\pm 0.059}$
0.24	$0.409_{\pm 0.012}$	$1.0_{\pm 0.0}$	$0.389_{\pm 0.047}$	$0.906_{\pm 0.034}$	$0.301_{\pm 0.015}$	$0.744_{\pm 0.092}$	$0.317_{\pm 0.047}$	$0.794_{\pm 0.041}$

Table 11: AM-Lag (top) and RbSL (bottom) baseline results for different levels of expert-random data mixture, retrained on each box width.

Expert %	**0**		**10**		**20**		**30**	
Box Width	MNC	SR	MNC	SR	MNC	SR	MNC	SR
0.16	0.011	0.95	0.013	0.949	0.01	0.932	0.021	0.941
0.20	0.255	0.955	0.218	0.945	0.142	0.941	0.146	0.935
0.24	0.457	0.964	0.405	0.926	0.33	0.879	0.323	0.833

Expert %	**0**		**10**		**20**		**30**	
Box Width	MNC	SR	MNC	SR	MNC	SR	MNC	SR
0.16	0.362	1.0	0.353	1.0	0.258	0.983	0.203	0.933
0.20	0.588	0.995	0.558	0.994	0.413	0.972	0.414	0.906
0.24	0.729	1.0	0.695	0.982	0.704	0.976	0.651	0.913

Table 12: Results for `UMazeObstacle` with alternative representation (avg. over 3 seeds).

# Avoid	**RADT**$^{\#}$		**AM-Lag**		**RADT-AltRep**$^{\#}$	
	MNC (1e-2)	SR	MNC (1e-2)	SR	MNC (1e-2)	SR
3	$4.455_{\pm 0.895}$	$0.868_{\pm 0.028}$	$4.47_{\pm 0.94}$	$0.645_{\pm 0.01}$	$4.587_{\pm 0.805}$	$0.91_{\pm 0.035}$
5	$8.75_{\pm 0.531}$	$0.853_{\pm 0.015}$	$10.14_{\pm 1.9}$	$0.76_{\pm 0.01}$	$8.897_{\pm 1.127}$	$0.883_{\pm 0.038}$
7	$12.7_{\pm 1.0}$	$0.832_{\pm 0.038}$	$12.72_{\pm 0.702}$	$0.777_{\pm 0.028}$	$12.723_{\pm 0.804}$	$0.887_{\pm 0.05}$

$^{\#}$ = Model capable of zero-shot generalization. Results highlighted in blue are zero-shot results.

Table 13: Results for the avoid-region relabeling ablation study on the `FetchReachObstacle` environment (avg. over 3 seeds).

Box Width	**+ avoid-region relabeling**		**- avoid-region relabeling**	
	MNC	SR	MNC	SR
0.16	$0.049_{\pm 0.016}$	$0.989_{\pm 0.010}$	$0.372_{\pm 0.023}$	$0.983_{\pm 0.016}$
0.18	$0.107_{\pm 0.018}$	$0.978_{\pm 0.009}$	$0.516_{\pm 0.06}$	$0.978_{\pm 0.01}$
0.2	$0.164_{\pm 0.027}$	$0.983_{\pm 0.017}$	$0.654_{\pm 0.006}$	$0.972_{\pm 0.025}$
0.22	$0.247_{\pm 0.018}$	$0.989_{\pm 0.019}$	$0.667_{\pm 0.07}$	$0.961_{\pm 0.01}$
0.24	$0.409_{\pm 0.012}$	$1.0_{\pm 0.0}$	$0.784_{\pm 0.009}$	$0.983_{\pm 0.017}$

Results highlighted in blue are zero-shot results.

Table 14: Results for the `MotionPlanning` environment that involves simultaneous varying of avoid-region number, location, shape, and size.

	RADT	**AM-Lag**	**RbSL**
MNC (1e-2)	**2.12**	3.81	5.14
SR	**0.99**	**1.0**	0.41

Table 15: Results for the `MazeObstacleImage-NumAvoid1` environment (avg. over 3 seeds).

	RADT	**AM-Lag**	**RbSL**	**WGCSL**
MNC (1e-2)	$\mathbf{1.1}_{\pm 0.95}$	$1.6_{\pm 1.23}$	$1.87_{\pm 0.78}$	$2.55_{\pm 1.44}$
SR	$\mathbf{0.617}_{\pm 0.029}$	$0.427_{\pm 0.031}$	$0.557_{\pm 0.121}$	$\mathbf{0.62}_{\pm 0.072}$

Table 16: Results for discrete versions of baselines on `Cardiogenesis`.

Scenario + setup	**discretized AM-Lag**			**dicretized RbSL**		
	% visitation	mean length	SR	% visitation	mean length	SR
Scenario 1 -avoid	1.0	29.4	0.14	1.0	4.2	1.0
Scenario 1 +avoid	0.0	31.1	0.0	0.0	31.0	0.0
Scenario 2 -avoid	0.51	13.17	0.67	0.32	18.29	0.59
Scenario 2 +avoid	0.33	15.68	0.69	0.26	4.53	1.0
Scenario 3 -avoid	1.0	27.06	0.28	1.0	11.6	0.98
Scenario 3 +avoid	1.0	29.32	0.12	0.0	26.6	0.0

Table 17: Failure rates for the `UMazeObstacle` (left) and `AMazeObstacle` (right) environments.

# Avoid	RADT$^{\#}$	AM-Lag	RbSL	# Avoid	RADT$^{\#}$	AM-Lag	RbSL
1	0.103	0.113	0.127	1	0.104	0.13	0.12
3	0.34	0.337	0.348	3	0.152	0.16	0.29
5	0.56	0.513	0.621	5	0.312	0.3	0.33
7	0.591	0.563	0.621	7	0.453	0.46	0.45

$^{\#}$ = Model capable of zero-shot generalization. Results highlighted in blue are zero-shot results.

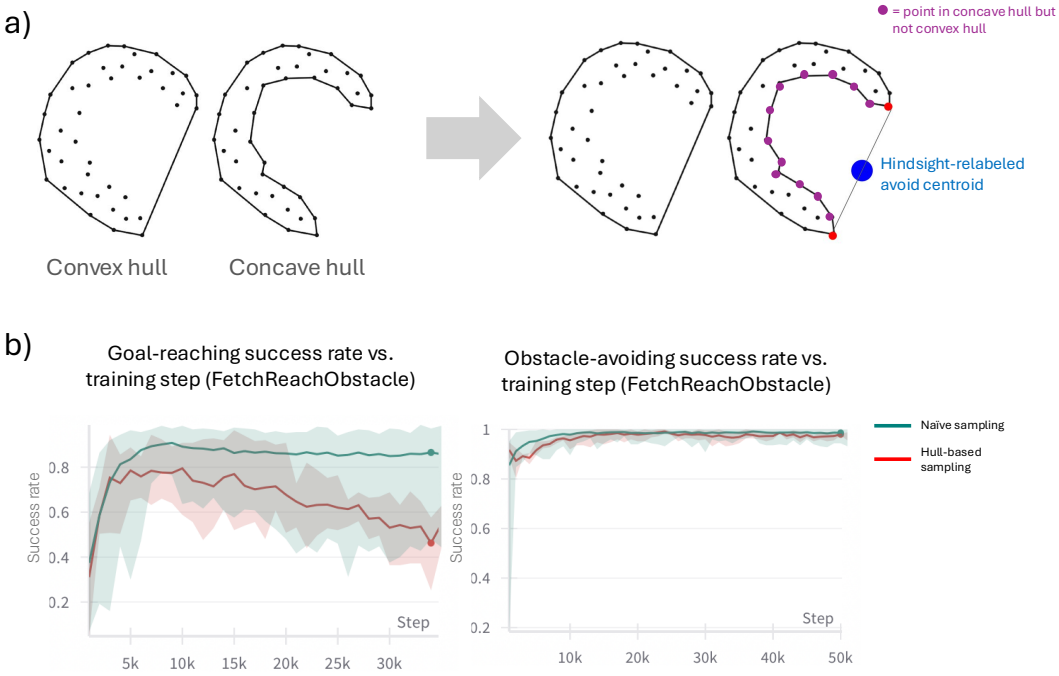


Figure 5: a) 2D depiction of the contour-based centroid sampling strategy for avoiding region relabeling. The convex and concave hulls are calculated for the set of data points in the state space for a training trajectory. Points that are part of the concave hull but not the convex hull (depicted here in purple) comprise concave portions of the trajectory that can be viewed as "wrapping around" some unknown avoid centroid. To generate an avoid centroid that fits this intuitive interpretation in hindsight, we locate the two points bordering this concave region (depicted here in red) and sample a point in between them. The resulting trajectory of data points looks like it's "trying" to avoid the blue hindsight-relabeled avoid centroid by wrapping around it. Figure built upon an illustration from [56]. b) Using the contour-based strategy for sampling avoids centroids results in a higher maximum SR for the `FetchReachObstacle` environment and notably mitigates/delays overfitting.

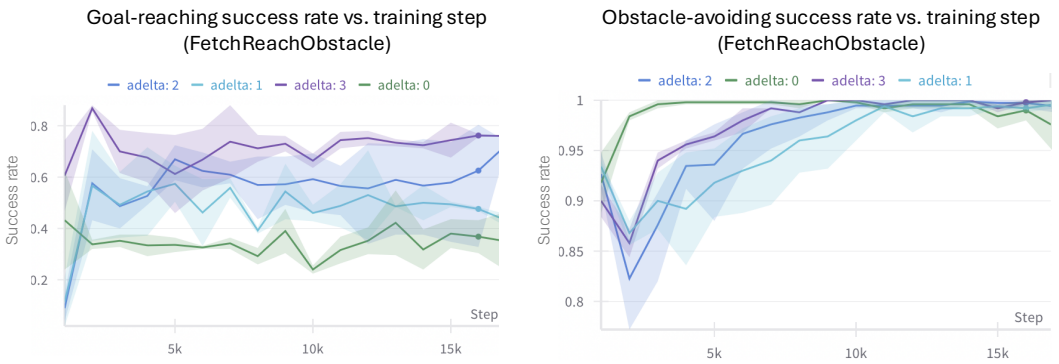


Figure 6: The maximum goal-reaching success rate for RADT trained on the FetchReachObstacle task improves with increasing the attention boosting bias `adelta` to the prompt tokens. The maximum obstacle-avoiding success rate, on the other hand, seems to be unaffected by the value of `adelta`.

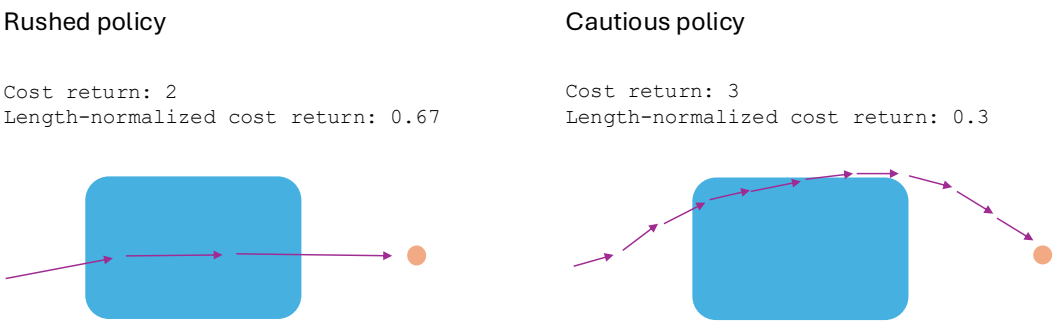


Figure 7: A policy that rushes directly through the avoid-region to get to the goal as quickly as possible may earn a low absolute cost return but a high length-normalized cost return. On the other hand, a slower, more cautious policy that takes more timesteps to reach the goal but attempts to circumvent the avoid-region may accumulate a similar absolute cost return, but a lower length-normalized cost return.

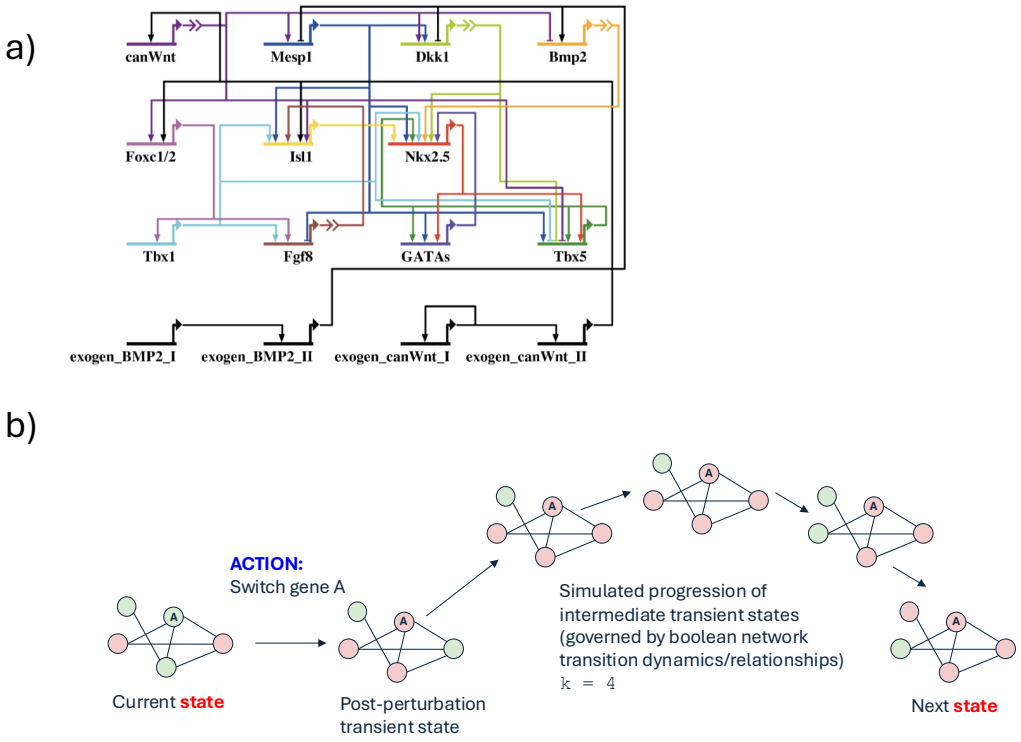


Figure 8: a) A diagram representing the 15-gene boolean network model for mouse cardiogenesis. b) A depiction of one state-action-state transition simulated in the `Cardiogenesis` environment, where `k = 4`.

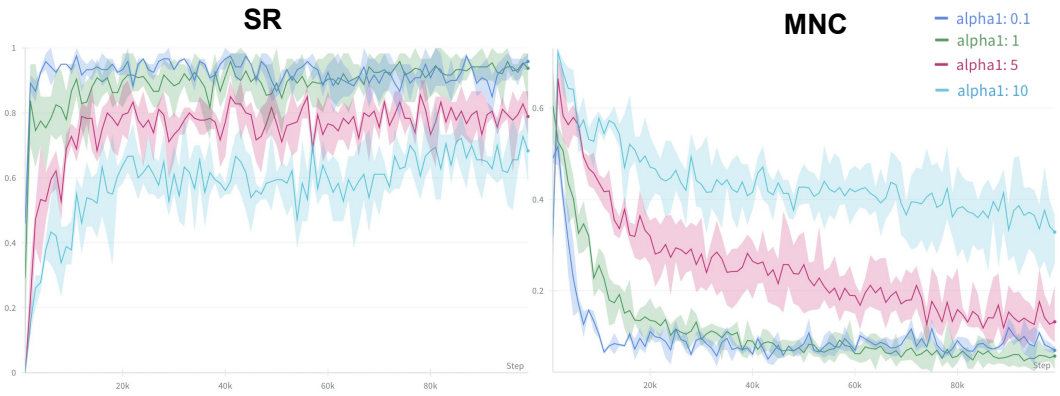


Figure 9: The effect of different values for the loss function hyperparameter α (ranging from $\alpha = 0.1$ to $\alpha = 10$). High values of α (e.g. 5, 10) de-prioritze the action loss too much and leads to poor goal-reaching (SR) and avoidance (MNC) behavior in the resulting policy. Small α values (e.g., 0.1, 1) lead to convergence on a policy with strong SR and MNC; however, when α is too small ($\alpha = 0.1$), the lowest MNC achieved is not as low as the lowest MNC achieved when $\alpha = 1$. We chose the value $\alpha = 1$. Training curves shown are averaged across 3 replicate runs per unique value of α.

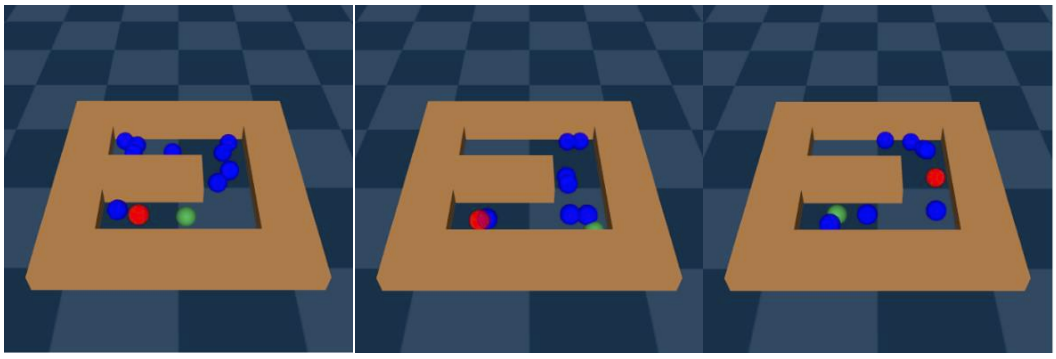


Figure 10: When the number of avoid-regions is high (e.g. 7 and 10 avoid-regions depicted in the U-Maze examples here), there is a often quite a lot of overlapping avoid-regions that compose together to form larger avoid-regions of more complex shape that the agent must circumvent.

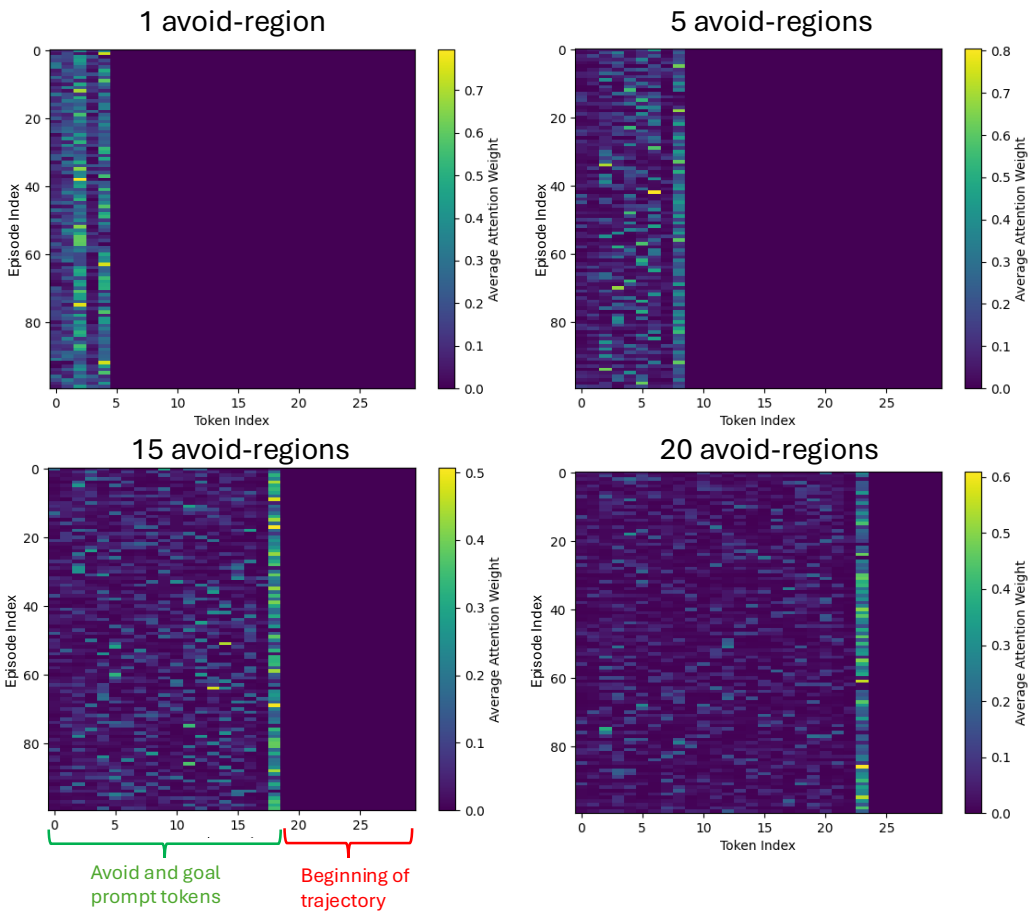


Figure 11: Plots of average attention weights to the first 30 tokens of the input sequence when RADT is outputting the last action of an evaluation episode (each episode is a row). The first 30 tokens include both the prompt (which include 1, 5, 15, or 20 avoid-region tokens) and the beginning of the trajectory history. RADT pays little attention to the beginning of the trajectory history as it has learned a mostly Markovian policy. No noticeable attention sinks and debilitating attention uniformity is observed with increasing prompt length, but the attention to individual avoid-region tokens on average does decrease with increasing prompt length. Attention to the goal token remains strong regardless of prompt length.

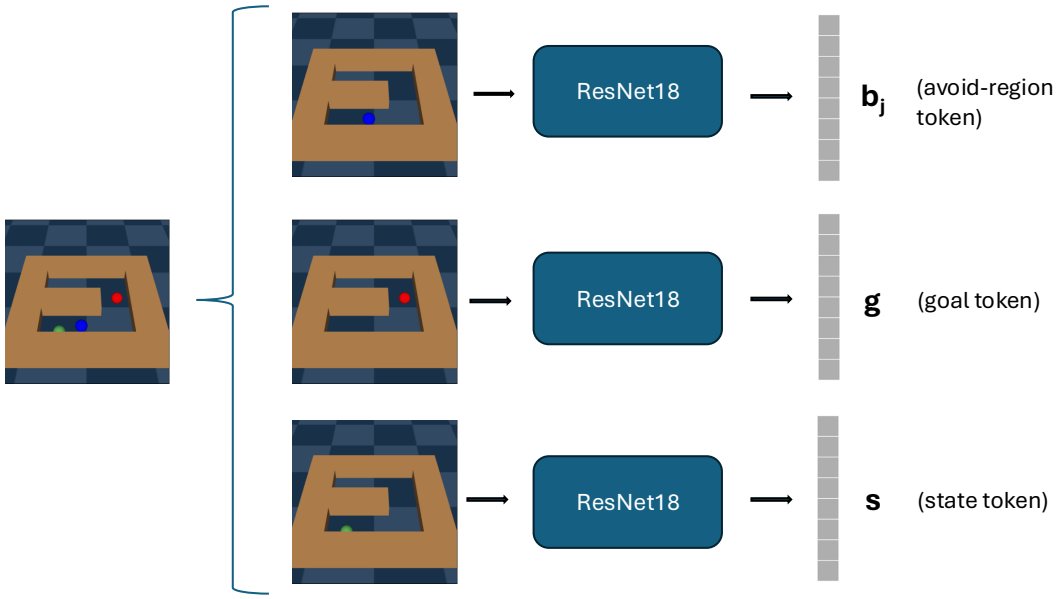


Figure 12: To obtain the avoid-region, goal, and state tokens for the `UMazeObstacleImage` environment, we use 224x224 images capturing the positions of the avoid-region(s) (blue), goal state (red), and agent position (green) in the maze. The images are fed through a pretrained ResNet-18 model to obtain 1000-dimensional flattened vector representations $\mathbf{b}_j$, $\mathbf{g}$, and $\mathbf{s}$.

J Additional References

Illustration Citations (NIAID NIH Bioart)

[The following images are included in our figures.]

NIAID Visual & Medical Arts., (10/7/2024). Fibroblast. NIAID NIH BIOART Source. bioart.niaid.nih.gov/bioart/ 152, a.

NIAID Visual & Medical Arts., (10/7/2024). Fibroblast. NIAID NIH BIOART Source. bioart.niaid.nih.gov/bioart/ 153, b.

NIAID Visual & Medical Arts., (10/7/2024). Fibroblast. NIAID NIH BIOART Source. bioart.niaid.nih.gov/bioart/ 154, c.

NIAID Visual & Medical Arts. (10/7/2024). Generic Im- mune Cell. NIAID NIH BIOART Source. bioart.niaid. nih.gov/bioart/173.

NIAID Visual & Medical Arts. (10/7/2024). Human Male Outline. NIAID NIH BIOART Source. bioart.niaid.nih. gov/bioart/232.

NIAID Visual & Medical Arts. (10/7/2024). Intermediate Progenitor Cell. NIAID NIH BIOART Source. bioart.niaid.nih.gov/bioart/258

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly listed 4 main contributions at the end of the Introduction. Contributions 1 and 2 were detailed in Section 4, contributions 3 and 4 were demonstrated in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Yes, we include a section on Limitations in Appendix G.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper has no rigorous theoretical results and does not claim any rigorous theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Hyperparameters and other technical specifications (e.g. RL environment specifications) are detailed in the Appendix. A GitHub repository is provided in Appendix H, with relevant commands to run main experiments in the README.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  4.1. If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  4.2. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  4.3. If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  4.4. We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Hyperparameters and other technical specifications (e.g. RL environment specifications) are detailed in the Appendix. A GitHub repository is provided in Appendix H, with relevant commands to run main experiments in the README.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters and other technical specifications (e.g. data prep, RL environment specifications) are detailed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We train each model on each benchmarking task 3 times, and include the standard deviation both in the table and as represented by error bars in the figures. This is standard practice for RL papers. No notion of statistical significance is defined.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute resources requirements are listed in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our data contains no information that can be used to harm individuals, as they are all produced from simulations. Our study does not involve human subjects. Our work does not seem to be at high risk for the potential concerns described in "Societal Impact and Potential Harmful Consequences".

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the Appendix, we include a section on Broader Impacts and discuss how our work provides foundational developments that can be potentially misused in downstream applications, including the potential misuse of autonomous vehicles/robots and bioengineered/genetically engineered cells.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All our data are generated from simulations and have no known risk of misuse. Our models are also trained to be run on simulated environments and have no clear risk of potential misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All original papers for baseline models are cited. All creators of benchmarking environments are cited. We built our repo off of MGPO and RbSL, and credited them in the README and Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We document our code in the provided anonymized repo.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We do not have human subjects involved in this study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: We do not have human subjects involved in this study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLMs were used for core method development.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.