

# Learning from Offline Heterogeneous Demonstrations via Reward-Policy Distillation Supplementary

Anonymous Author(s)

Affiliation

Address

email

## 1 Offline LfD Enhancements Detail

AVRIL considers a distribution over the reward function and approximates the posterior,  $p(R|\mathcal{D})$ , with a variational distribution,  $q_\phi(R)$ . It is trained by maximizing the Evidence Lower Bound (ELBO), shown in Equation 1, where  $p(R)$  is the prior distribution for the reward function,  $\pi_E$  is the expert policy. The second equation follows by the assumption of Boltzmann rationality of the demonstrator [1].

$$\begin{aligned} \text{ELBO}(\phi) &= \mathbb{E}_{q_\phi} [\log p(\mathcal{D}|R)] - D_{\text{KL}}(q_\phi(R)||p(R)) \\ &= \mathbb{E}_{q_\phi} \left[ \sum_{(s,a) \in \mathcal{D}} \log \frac{\exp(\beta Q_R^{\pi_E}(s,a))}{\sum_{b \in \mathbb{A}} \exp(\beta Q_R^{\pi_E}(s,b))} \right] - D_{\text{KL}}(q_\phi(R)||p(R)) \end{aligned} \quad (1)$$

Directly optimizing the ELBO is not feasible as the gradient of  $Q_R^{\pi_E}(s,a)$  with respect to  $\phi$  is intractable. Therefore, AVRIL introduces a second variational approximation for  $Q_R^{\pi_E}(s,a)$  with  $Q_\theta(s,a)$  and ensures the variational reward distribution,  $q_\phi(R)$ , is consistent with the variational Q function,  $Q_\theta(s,a)$ , by Bellman equation, i.e.,  $R(s,a) = \mathbb{E}_{s',a' \sim \pi} [Q_R^\pi(s,a) - \gamma Q_R^\pi(s',a')]$ .

Here, we present a lemma to show how AVRIL Enhancement 1 (i.e., extending KL-divergence regularization on all actions) impacts reward learning. This enhancement could be viewed as a data augmentation technique to encourage a small distance to the prior distribution for any action for each state in the demonstration. We formalize the intuition in Lemma 1.

**Lemma 1.** Assume the prior reward distribution,  $p(R(s,a))$ , is a Gaussian distribution partitioned on each state and action pair; minimizing  $L_{\text{KL}}$  results in  $q_\phi(R(s,a)) = p(R(s,a))$  for each operated  $(s,a)$ .

Following Lemma 1 and our extended operation over  $b \in \mathbb{A}$ , we have the following observation.

**Corollary 1.1.** Assume we choose the prior reward distribution  $p(R(s,a))$  to be Standard Gaussian distribution,  $\mathcal{N}(0,1)$ . for  $s \in \mathcal{D}, b \in \mathbb{A}$  s.t.  $(s,b) \notin \mathcal{D}$ , optimizing  $L_{\text{AVRIL}}$  leads to  $\mu_\phi(s,b) = 0$  and  $\sigma_\phi^2(s,b) = 1$ . The proof follows immediately by observing  $q_\phi(R(s,b))$  only gets gradient from  $L_{\text{KL}}$  and the optimal solution of  $L_{\text{KL}}$  is that  $\mu_\phi(s,a) = 0$  and  $\sigma_\phi^2(s,a) = 1$ .

## 2 MPP Heterogeneity Analysis Details

In our analysis, we seek to compare the variance of path features within each RP to the variance of path features across RPs, as this would help quantify how diverse expert demonstrations are. The demonstration for each RP is multivariate and not normal (after performing tests for normality and homoscedasticity), necessitating the use of the PERMANOVA test, which is non-parametric and is able to compare multivariate data. More specifically, it tests the null hypothesis that the centroid and dispersion for two groups are equivalent. To apply this test to the RP data, we tested each possible pair of RPs to see which ones have statistically significant differences in their distribution.

Hyperparameters	Values
Training Itrs	1000
Learning Rate	0.0001
State Only Reward	False
State Dim	4, 9
Action Dim	2
Gamma	0.99
Lambda	1.
Train Test Split	0.8
Min Number of Test Sols	1
Linear Reward	False
Offline CQL Training Itrs	1000
Strategy Reward Regularization Coefficient (MSRD and DROID)	0.01
Strategy Q Function Regularization Coefficient (DROID)	0.001

Table 1: This table shows the hyperparameters we use for DROID and all benchmark algorithms. All values separated with commas are for CartPole and MPP, respectively.

Therefore, if the PERMANOVA test has a low p-value for two RPs, this indicates that there is a significant amount of variation between those two RPs compared to within those two RPs. The Bonferroni-Holm method was used to account for the fact that many hypothesis tests are being performed.

### 3 Experimental Setup

For fair comparison on all baseline techniques, we share the same network architecture for each policy and reward with two hidden layers of 64 units along with GELU activation functions trained using Adam for 1000 iterations. For downstream policies, we train offline Conservative Q Learning [2] with several improvements proposed in Rainbow [3] for 1000 iterations. Conservative Q-learning is an offline RL algorithm designed to guard against overestimation while avoiding explicit construction of a separate behavior model. We leverage several improvements including Dueling Double Q Networks and Distributional RL from Rainbow [3] to improve the CQL training. We list hyperparameters used in all algorithms in Table 1.

In order to showcase the significance of our results on both the Cartpole and Mars datasets, we perform tests for normality and homoscedasticity and find that our metrics do not satisfy the assumptions of a parametric ANOVA test. Thus, we instead perform a non-parametric Friedman test followed by a posthoc Nemenyi–Damico–Wolfe (Nemenyi) test. We show significance by aligning subject groups (demonstrations) between different treatments (benchmark techniques) along each demonstration.

#### 3.1 CartPole

##### 3.1.1 Video Demonstrations

We include demonstrations of heterogeneous behaviors along with each technique’s learned policies in CartPole in the link: <https://tinyurl.com/droidcartpolevideos>.

##### 3.1.2 Metrics

Here, we describe the motivation behind each of the metrics, evaluated from rollouts of the policies with respect to expert demonstrations.

1. Frechet Distance [4]: Compare the spatial and temporal differences of the trajectory from the agent’s policy with the expert trajectory to quantify how well the agent captures the motion pattern of the expert.

2. KL Divergence [5]: By estimating the state distribution within a trajectory by the kernel density estimator [6], this quantifies how well the learned policies state visitation matches the expert’s.
3. Undirected Hausdorff Distance [7]: This measures the maxima between the two Directed Hausdorff distances: one mapping our learned policy’s trajectory to the expert trajectory, and the other mapping the expert trajectory to our learn policy’s trajectory. This metric studies how far the agent’s trajectory is from the expert’s trajectory.
4. Average Log Likelihood: This measures the likelihood of expert demonstration under the learned policy.

### 3.1.3 Analysis

We showcase metrics in which DROID outperforms the baseline techniques in CartPole for the three experiments (Diverse Demonstration Modeling, Policy Transferability, and Reward Generalizability, c.f. main paper Result Section Q1-Q3) in Figure 1 and the statistics in Table 2. In the training task, DROID generates rollouts that align closer with expert behaviors, evidenced by stronger Undirected Hausdorff performance. Likewise, DROID does significantly better on ”Log Likelihood” on all tasks compared to the best baseline. Common reward-policy distillation helps guide DROID’s policies and rewards to better modeling expert preferences and thus better captures diversity in expert behaviors.

## 3.2 Mars Path Planning

### 3.2.1 Domain Introduction

Exploring Mars has been a fascinating and challenging endeavor for space agencies around the world. The Curiosity Rover is the longest active autonomous vehicle NASA has sent to Mars to study the climate, geology, and potential habitability of the planet [8]. The Rover has been in operation for the past ten years and its path planning has been done by manual labor of Rover Planners (RPs) on Earth.

There are several factors RPs consider when designing paths, including the change in elevation, distance to the desired destination, uncertainty about missing data on the terrain, etc. We study a dataset of Curiosity Rovers curated paths from 163 sols (a sol being a Martian day, approximately 24.6 hours). We demonstrate in Section 4 of the main paper that there is significant heterogeneity

Table 2: This table shows the APA-style statistical test results for Friedman ( $\alpha = 0.05$ , d.o.f.=3), Posthoc Nemenyi ( $\alpha = 0.05$ ) of DROID with respect to baselines in Cartpole (left). All reported test statistics are significant other than the italicized metrics where no posthoc analysis is performed.

CartPole				
Benchmark Method	KL Divergence	Frechet Distance	Undirected Hausdorff	Log Likelihood
<b>Diverse Demonstration Modeling (<math>n = 40</math>)</b>				
Friedman	26.31	49.26	50.76	104.25
DROID vs AVRIL Batch	<i>1.95</i>	<i>2.42</i>	3.11	-3.46
DROID vs AVRIL Single	3.81	4.33	3.63	-5.63
DROID vs MSRD	4.24	6.75	1.10	1.95
<b>Policy Transferability (<math>n = 20</math>)</b>				
Friedman	14.10	25.44	28.5	51.54
DROID vs AVRIL Batch	<i>0.86</i>	<i>0.45</i>	3.18	2.81
DROID vs AVRIL Single	<i>0.97</i>	<i>0.46</i>	<i>0.32</i>	3.80
DROID vs MSRD	3.55	4.89	5.26	7.10
<b>Reward (<math>n = 20</math>)</b>				
Friedman	2.22	<i>0.54</i>	<i>0.30</i>	13.38
DROID vs AVRIL Batch	N/A	N/A	N/A	2.81
DROID vs AVRIL Single	N/A	N/A	N/A	2.08
DROID vs MSRD	N/A	N/A	N/A	3.43

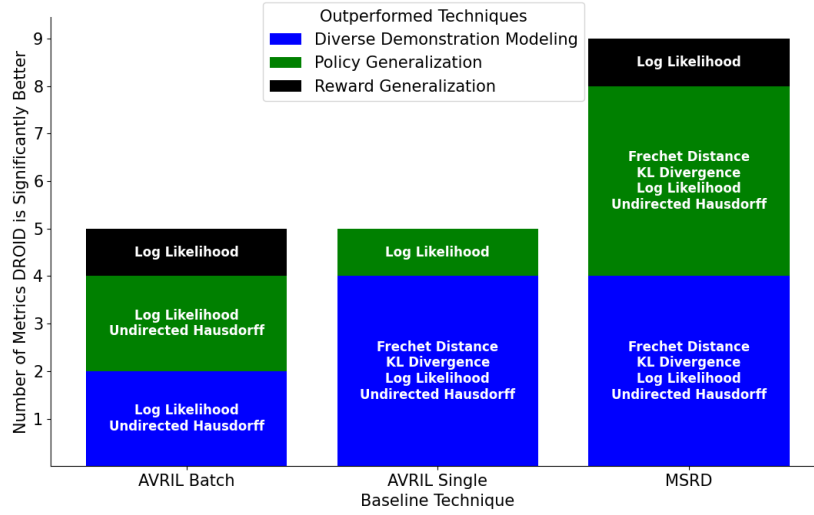


Figure 1: This barchart indicates along which metrics DROID outperforms the baseline techniques in offline imitation learning, generalization, and reward transfer tasks using a posthoc Nemenyi-Friedman analysis paired along each Sol.

88 between RP’s paths. Each RP has a specific priority among safety, efficiency, risk, and mission  
 89 constraints that inform their path design. This motivates us to design an autonomous path-planning  
 90 approach that learns from these heterogeneous experts.

### 91 3.2.2 Dataset Curation

92 The data consists of features that were created through a series of interviews with RPs, scientists,  
 93 and engineers that ideally capture the decision-making process for rover path planning. The features  
 94 were engineered to codify the reasoning behind the RP’s decisions. For example, RPs would visually  
 95 analyze the terrain and map waypoints to avoid “rough” terrains but without quantifiable measures  
 96 of what is considered rough. We identify the following features to encode the mental models and  
 97 strategies of RPs with considerations of risks, efficiency, safety, and mission requirements.

Table 3: This table shows the APA-style statistical test results for Friedman ( $\alpha = 0.05$ , d.o.f.=3), Posthoc Nemenyi ( $\alpha = 0.05$ ) of DROID with respect to baselines in MPP. All reported test statistics are significant other than the italicized metrics where no posthoc analysis is performed

Mars Path Planning				
Benchmark Methods	Undirected Hausdorff	Distance from Waypoint	Final Distance	Log Likelihood
<b>Diverse Demonstration Modeling (<math>n = 114</math>)</b>				
Friedman	153.94	71.58	223.59	29.74
DROID vs AVRIL Batch	2.82	2.98	2.54	4.61
DROID vs AVRIL Single	3.10	1.33	11.95	4.36
DROID vs MSRD	2.79	2.31	3.74	14.62
<b>Policy Transferability (<math>n = 49</math>)</b>				
Friedman	41.29	19.53	68.98	83.59
DROID vs AVRIL Batch	4.39	4.99	3.35	<i>1.56</i>
DROID vs AVRIL Single	4.70	<i>2.15</i>	5.05	<i>0.78</i>
DROID vs MSRD	5.08	<i>2.14</i>	6.41	6.57
<b>Reward (<math>n = 49</math>)</b>				
Friedman	10.81	<i>0.47</i>	78.68	<i>0.77</i>
DROID vs AVRIL Batch	3.48	N/A	5.91	N/A
DROID vs AVRIL Single	3.24	N/A	6.10	N/A
DROID vs MSRD	2.72	N/A	6.03	N/A

98 **Distance Feature** – The distance feature measures the percent added distance the rover must take  
 99 with the addition of intermediate waypoints in relation to the direct distance between the start and  
 100 end waypoints. The aim of this feature is to drive the rover’s necessary additional distance. It  
 101 is assumed that the path between waypoints is driven straight as RPs rarely drive curved paths  
 102 and rather set additional intermediate points in the event the rover needs to avoid hazards between  
 103 waypoints.

104 **Unknown Data Feature** – In the construction of the height maps, data could be missing where the  
 105 cameras are unable to see terrain beyond a hill or obstructions like a large rock or the rover itself.  
 106 By traversing terrains with missing data, the RP places the rover at a higher risk of being damaged.  
 107 The design of the Unknown Data Feature is to minimize the distance the rover traverses over terrain  
 108 without data. We compute the unknown data feature as the percent data missing in the height map  
 109 for proposed trajectories.

110 **Roughness Feature** – Rover Planners ideally drive on relatively smooth surfaces, avoiding rough  
 111 terrain that could potentially damage the rover’s hardware. Similarly, Rover Planners also look to  
 112 avoid terrain that is too soft, to prevent a similar fate as Spirit getting stuck in sand [9]. Here,  
 113 the roughness is computed as the difference of consecutive surface angles as the rover traverses to  
 114 the goal point. The maximum roughness and the average roughness over proposed trajectories are  
 115 measured to avoid large holes or rocks as well as minimize traveling on rough terrains.

116 **Pitch and Roll Feature** – Pitch and roll of the rover’s orientation adds another level of safety checks  
 117 that ensures that the rover will not face terrains that risk the rover rolling over.

118 **Turning Trajectory** – We include the turning trajectory as a feature to track. This feature calculates  
 119 the angle the rover must turn at intermediate waypoints. With this feature, the cost of taking sharp  
 120 turns considers the rover’s hardware and long-term health.

**Waypoint Grid Construction** – The 64x64 sized waypoint grid is constructed by scaling the terrain  
 height map along each axis and sampling the terrain map height at each (x,y) coordinate in the scaled  
 grid. We do so according to an inverse weighted distance from each point along the 4 nearest points  
 with height map data. We perform this scaling to limit the size of the action space of possible  
 waypoints we can visit.

$$H(x, y) = \frac{\frac{h_1}{d_1} + \frac{h_2}{d_2} + \frac{h_3}{d_3} + \frac{h_4}{d_4}}{\frac{1}{d_1} + \frac{1}{d_2} + \frac{1}{d_3} + \frac{1}{d_4}}$$

121  $H$  represents the height evaluated at each point in the gaming area.  $(x, y)$  represent the correspond-  
 122 ing coordinates and  $h_i, d_i$  represent the height and distance away from the evaluated point in the  
 123 dataset respectively.

### 124 3.2.3 Description of Policy

125 The action space exists on a 64 by 64 discrete grid of 4096 possible successor waypoints. This  
 126 was chosen to be large enough so that we can have high precision when selecting waypoints. The  
 127 average distance between grid points is ranges from 0.01m to 0.7m. We define our learned policy  
 128 in Equation 2 from our learned Q-function  $Q_\theta$ .

$$\pi_\theta(s) = \max_{a \in A} Q_\theta(s, a) \quad (2)$$

129 As mentioned in Section Preprocessing, we consider the three-waypoint planning problem and there-  
 130 fore, an action,  $a$ , (i.e., the intermediate waypoint) determines the trajectory as from the current point  
 131 to the intermediate waypoint and then from the intermediate waypoint to the ending waypoint. We  
 132 calculate the features of the action (i.e., next waypoint) for each of the two segments of the trajectory  
 133 (current point to next waypoint, and next waypoint to goal point).

### 134 3.2.4 Metrics

135 Here we include further description of the metrics we study in the MPP problem:

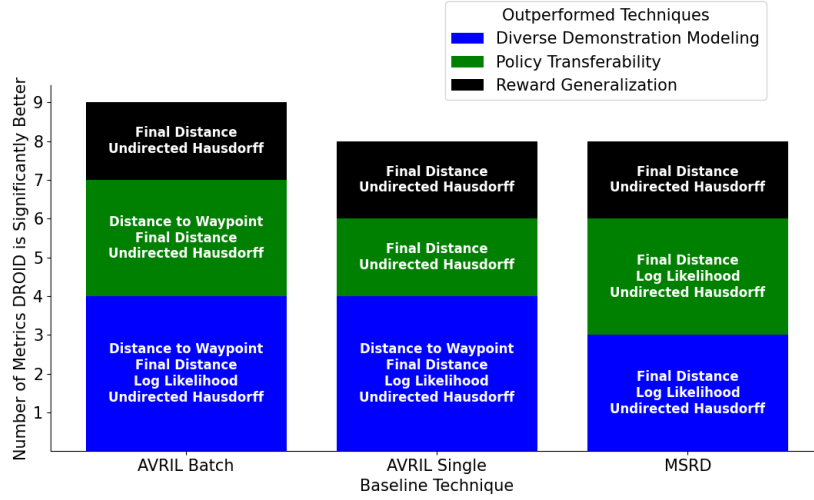


Figure 2: This barchart indicates along which metrics DROID outperforms the baseline techniques on our training, testing, and downstream reward transfer tasks using a posthoc Nemenyi-Friedman analysis paired along each Sol.

1. Average Distance from Midpoint: The average distances from our policies' predicted waypoints to the demonstrated waypoints.
2. Distance from Endpoint: The average distance from the final waypoint selected by the path generated by each technique to the goal point.
3. Undirected Hausdorff Distance [7]: This metric measures the maxima between the Directed Hausdorff distances mapping both our learned policy's set of waypoints to the expert waypoints and vice-versa.
4. Average Log Likelihood: This metric measures the likelihood of expert demonstration under the learned policy.

### 3.2.5 Analysis

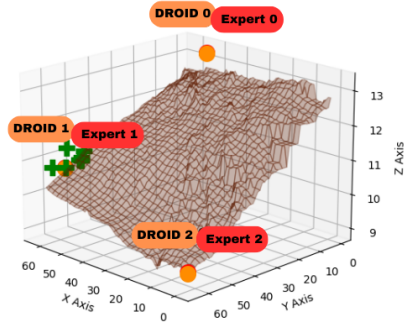
We showcase the specific metrics that DROID outperforms baseline techniques on for MPP in Figure 2 and statistics in Table 3. Rather than assuming homogeneity across demonstrations or discarding data to design a personalized policy for each RP, DROID takes advantage of per-RP modeling and knowledge sharing to significantly outperform on 3/4 metrics in Offline IRL benchmark. On the policy generalization benchmark, DROID is also able to model the latent objectives from diverse experts to induce a trajectory in unseen Sols that align closer to the expert's true path while successfully capturing the high-level common task goal. Lastly, DROID is the only technique to show significantly better performance on downstream reward transfer indicating the learned reward is a more useful encoding of an expert's latent objective and can be used to better interpret the salient features for a given expert.

### 3.2.6 Additional Qualitative Analysis

In this section, we discuss the additional contributions of DROID to the goal of interpreting expert decision-making and how it is valuable in the domain of path planning for the Mars Curiosity Rover and future missions.

**Decomposition Learned Policy** First, we analyze the our learned Q-function and study how it can provide insight into the decision-making process of our model. We showcase in Figure 3 how our technique can highlight the top 10 highest-rated successor waypoints from the start position and

Normalized 3D Terrain Map of Top Waypoints from Start in Sol 2575



Normalized 3D Terrain Map of Top Waypoints from Waypoint 1 in Sol 2575

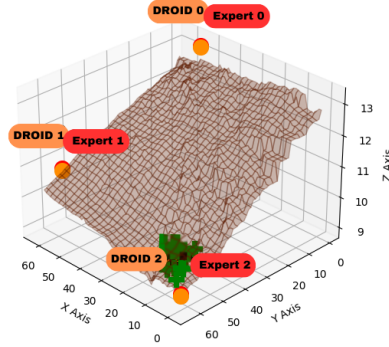


Figure 3: These figures show DROID’s policy outputs on a terrain map to plan for the next waypoint from the Start point (left) and Waypoint 1 (right). The orange spheres represent the selected waypoints of DROID and the expert. Highlighted in green above the terrain map are the top 10 highest-rated successor waypoints. The orange labels correspond to DROID’s found waypoints and the red labels correspond to the expert demonstration’s waypoints.

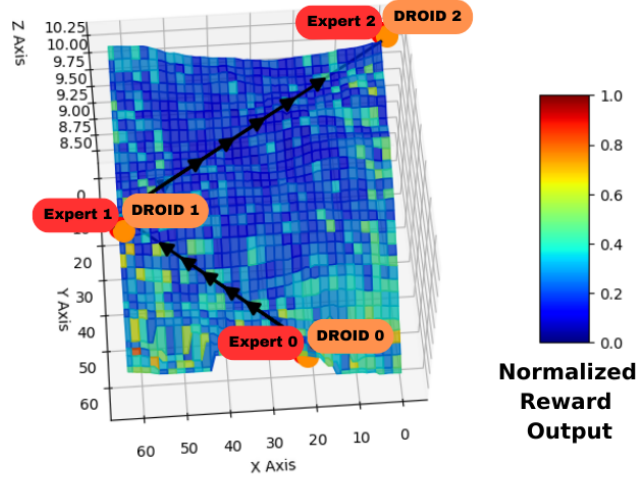


Figure 4: This figure shows a heatmap of Sol 2163 of DROID’s Strategy Reward log standard deviation estimate, where higher value represents greater uncertainty in the reward estimation.

163 midpoint position respectively. Providing multiple options aligned with the expert’s latent prefer-  
 164 ence could be beneficial for a future assistive tool for RPs. Expert drivers at NASA can also study  
 165 DROID’s recommended waypoint and similar waypoints that are rated highly by our model. If an  
 166 expert disagrees with the best action identified by DROID, we can find several additional options  
 167 that align with that expert’s latent preferences.

168 **Uncertainty of Reward Estimate** We examine the uncertainty in the reward predictions of our  
 169 model by plotting the standard deviation of the strategy reward posterior. As shown in Figure 4,  
 170 we can estimate how uncertain our model is about different parts of the terrain due to the limited  
 171 coverage of the dataset. Intuitively, areas of the state space where the demonstrations have not  
 172 covered, such as the edges of the terrain, have a higher estimate of uncertainty.

173 **Proposed Application to NASA** The explainability provided by DROID has significant potential  
 174 application as a supplementary planning tool for interplanetary exploration. By leveraging a Shapely  
 175 value analysis of the importance of different factors such as the change in elevation and uncertainty

176 about missing data on the terrain, rover planners can gain a deeper understanding of the objective  
177 function our algorithm extracts for modeling rover path planning. Therefore, DROID can explain  
178 what features contribute most to its perceived estimate of any human or AI-designed path.

179 This may have application to informing the design of future missions by providing more insight  
180 into the limitations of the rover and the types of environments in which it is best suited to operate.  
181 As shown on the Curiosity Rover dataset, DROID could be applied to give rapid feedback about  
182 paths that best avoid sharp rocks (Rough Terrain) which may damage the open holes on the rover’s  
183 wheels [10], thus improving the longevity of the rover. Similarly, with additional data, such as  
184 orbital satellite imagery, our approach could be used to evaluate the value of landing sites [11] by  
185 studying our learned RP objective function on constructed terrain maps.

186 Furthermore, the ability to model different strategies taken by human drivers could potentially be  
187 used in the future by JPL in the development of training programs. We hope a future application  
188 of DROID would be to capture difficult-to-articulate tribal knowledge among rover planners and  
189 identify the most important features to trainees. We can describe implicitly understood knowledge  
190 to help train new drivers at NASA faster and with greater efficiency. By letting DROID explain  
191 which features have the greatest contribution to the underlying latent RP strategy, human drivers can  
192 better understand what features to consider when navigating other extraterrestrial terrains.

193 Our hope is that DROID lessens the burden for operators to plan out daily schedules for rovers (since  
194 it performs automated path planning that better optimizes for operator preferences). Moreover, the  
195 algorithm’s ability to reason under uncertainty makes it particularly useful for fast path planning in-  
196 ference, even when there is occluded information from cameras or other sensors. With the DROID’s  
197 ability to learn diverse expert strategies and plan under uncertainty/occlusions, our algorithm could  
198 further advance fast autonomous rover exploration.

### 199 3.2.7 Ablation

200 In this section, we perform an ablation study to evaluate the effectiveness of different components  
201 of our approach compared to our method DROID. Ablation 1-6 corresponds to the following:

- 202 1. DROID without AVRIL improvements 1 and 2
- 203 2. DROID without AVRIL Improvement 1
- 204 3. DROID without AVRIL Improvement 2
- 205 4. DROID without any distillation
- 206 5. DROID without policy distillation
- 207 6. DROID without reward distillation

208 Table 4 summarizes the results of our ablation study in the MPP problem. We find that on the Diverse  
209 Demonstration modeling task, DROID consistently outperforms all ablation variants across all met-  
210 rics (other than Distance from Waypoint which where it comes second), indicating the importance of  
211 both reward and policy distillation in our approach. We also observe that DROID without “AVRIL  
212 Improvement 1” and DROID without “AVRIL Improvement 2” perform worse than DROID, sug-  
213 gesting that both improvements are effective in improving the performance of our approach. We  
214 also find that removing policy distillation or reward distillation leads to a significant decrease in  
215 performance, indicating the importance of both types of distillation in our approach.

216 Likewise, table 5 showcases the results of the ablation study in the policy transferability task.  
217 “DROID without policy distillation” (Ablation 5)’s strong waypoint modeling performance is  
218 matched by DROID performs comparably and it outperforms Ablation 5 with much stronger per-  
219 formance on goal point finding. “DROID without any distillation” achieves strong Log Likelihood  
220 performance but does not match DROID’s generalization capacity where DROID has better Undi-  
221 rected Hausdorff, suggesting it can model expert’s desired waypoints closer than any other ablated  
222 techniques.

Table 4: This table shows the ablation performance of DROID along the Diverse Demonstration Modeling task. Bold indicates the best-performing model of the metric.

Method	Distance from Waypoint	Final Distance	Undirected Hausdorff	Log Likelihood
Ablation 1	4.871	1.557	8.391	-10.157
Ablation 2	6.084	0.288	7.575	-10.104
Ablation 3	7.126	0.571	7.287	-8.431
Ablation 4	6.720	0.209	7.441	-8.419
Ablation 5	<b>3.910</b>	4.014	7.498	-225.212
Ablation 6	5.556	6.783	9.389	-14.479
DROID	4.592	<b>0.070</b>	<b>6.780</b>	<b>-7.261</b>

Table 5: This table shows the ablation performance of DROID along the Policy Transfer task. Bold indicates the best-performing model of the metric.

Method	Distance from Waypoint	Final Distance	Undirected Hausdorff	Log Likelihood
Ablation 1	8.086	1.842	8.945	-15.010
Ablation 2	8.071	1.615	8.331	-16.334
Ablation 3	9.318	3.933	9.644	-16.503
Ablation 4	8.518	0.576	7.744	<b>-11.391</b>
Ablation 5	6.162	7.295	9.537	-13.610
Ablation 6	8.026	9.078	9.246	-30.037
DROID	<b>6.144</b>	<b>0.277</b>	<b>6.407</b>	-18.483

Overall, our ablation study confirms the effectiveness of our proposed approach and highlights the importance of both reward and policy distillation, as well as the two AVRIL improvements, in achieving state-of-the-art performance.

## 4 Additional Related Works

In this section, we describe additional related works regarding offline path planning under uncertainty and navigation beyond an MDP setting.

**Path Planning Algorithm.** Several works use human-inspired admissible heuristic functions to plan paths [12, 13]. Yet, these functions are handcrafted and require domain expertise to design. Model Predictive Path Integral (MPPI) is studied for local path following for rovers [14]. However, classical path planning approaches fail without a high-fidelity simulator [15, 16]. Other works look at the problem of path planning to maximize a reward function [17, 18, 19] under uncertainty. However, these techniques leverage exploration to obtain a better estimate of their cost function, which may not be feasible in offline learning. Our algorithm, DROID, learns heterogeneous preferences and policies directly from expert demonstrations, without the assumption of a hand-designed reward function or a simulator.

**Generalization Performance of Navigation Algorithms.** Another important factor in offline path planning is the generalization performance of the planning algorithm to novel terrains. To improve the generalization performance, existing work attempts to decouple the training of a feature extraction block and navigation block using Deep RL [20]. However, they perform online planning along 2D navigation cost-map to extract an attention map which is not feasible in the offline setting. Additionally, Meng et al. [21] proposes a path planning algorithm that balances the trade-off between safety and efficiency under uncertainty. However, this approach does not generalize to unseen environments that contain new or additional obstacles.

**Path Planning on the Martian Domain.** There are several prior works that study path planning in the Martian domain but focus on local path planning and do not address long path planning. Hedrick et al. [22] proposes efficient Martian path planning and Rover-IRL [11] learns a cost function from demonstration but both fail to plan under uncertainty which is a key assumption in the Mars domain [10].

## References

- [1] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the National Conference on Artificial intelligence (AAAI)*, pages 1433–1438, 2008.
- [2] A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *CoRR*, abs/2006.04779, 2020. URL <https://arxiv.org/abs/2006.04779>.
- [3] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. G. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. *CoRR*, abs/1710.02298, 2017. URL <http://arxiv.org/abs/1710.02298>.
- [4] K. Toohey and M. Duckham. Trajectory similarity measures. *SIGSPATIAL Special*, 7:43–50, 05 2015. doi:10.1145/2782759.2782767.
- [5] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.
- [6] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Probl. Inf. Transm.*, 23(1-2):95–101, 1987.
- [7] F. Hausdorff. *Grundzüge der Mengenlehre*. Chelsea, 1914. <https://en.wikipedia.org/wiki/Grundz>
- [8] A. R. Vasavada. Mission Overview and Scientific Contributions from the Mars Science Laboratory Curiosity Rover After Eight Years of Surface Operations. *Space Sci Rev*, 218(3):14, 2022.
- [9] W. Brown. Nasa to begin attempts to free sand-trapped mars rover, Nov 2009. URL [https://www.nasa.gov/mission\\_pages/mer/news/mer20091112.html](https://www.nasa.gov/mission_pages/mer/news/mer20091112.html).
- [10] D. M. Gaines, R. C. Anderson, G. B. Doran, W. Huffman, H. Justice, R. M. Mackey, G. R. Rabideau, A. R. Vasavada, V. Verma, T. A. Estlin, L. M. Fesq, M. D. Ingham, M. W. Maimone, and I. A. D. Nesnas. Productivity challenges for mars rover operations, 2016.
- [11] M. Pflueger, A. Agha, and G. S. Sukhatme. Rover-irl: Inverse reinforcement learning with soft value iteration networks for planetary rover path planning. *IEEE Robotics and Automation Letters*, 4(2):1387–1394, 2019. doi:10.1109/LRA.2019.2895892.
- [12] W. Gong, X. Xie, and Y.-J. Liu. Human experience-inspired path planning for robots. *International Journal of Advanced Robotic Systems*, 15:172988141875704, 02 2018. doi:10.1177/1729881418757046.
- [13] R. Ramón-Vigo, N. Pérez-Higueras, F. Caballero, and L. Merino. Transferring human navigation behaviors into a robot local planner. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 774–779, 2014. doi:10.1109/ROMAN.2014.6926347.
- [14] S. Dergachev, K. Muravyev, and K. Yakovlev. 2.5d mapping, pathfinding and path following for navigation of a differential drive robot in uneven terrain, 2022.
- [15] B. Hao, H. Du, J. Zhao, J. Zhang, and Q. Wang. A Path-Planning approach based on potential and dynamic Q-Learning for mobile robots in unknown environment. *Comput Intell Neurosci*, 2022:2540546, June 2022.
- [16] E. Murphy. Planning and exploring under uncertainty, Jan 2010. URL <https://ora.ox.ac.uk/objects/uuid3Abb3d85f6-117b-4f5e-92ab-b6acc87aef79>.
- [17] T. Yu, B. Deng, J. Gui, X. Zhu, and W. Yao. Efficient informative path planning via normalized utility in unknown environments exploration. *Sensors (Basel, Switzerland)*, 22, 2022.

- 295 [18] L. Cuevas, M. Ramírez, I. Shames, and C. Manzie. Path planning under risk and uncertainty  
296 of the environment. *2021 American Control Conference (ACC)*, pages 4231–4236, 2021.
- 297 [19] Y. Yin, Z. Chen, G. Liu, and J. Guo. A mapless local path planning approach using deep  
298 reinforcement learning framework. *Sensors*, 23:2036, 02 2023. doi:10.3390/s23042036.
- 299 [20] K. Weerakoon, A. J. Sathyamoorthy, U. Patel, and D. Manocha. Terp: Reliable planning in  
300 uneven outdoor environments using deep reinforcement learning, 2021.
- 301 [21] F. Meng, L. Chen, H. Ma, J. Wang, and M. Q.-H. Meng. Nr-rrt: Neural risk-aware near-optimal  
302 path planning in uncertain nonconvex environments. *arXiv preprint arXiv: 2205.06951*, 2022.
- 303 [22] G. Hedrick, N. Ohi, and Y. Gu. Terrain-aware path planning and map update for mars sample  
304 return mission. *IEEE Robotics and Automation Letters*, 5(4):5181–5188, 2020. doi:10.1109/  
305 LRA.2020.3005123.