Interactive Terrain Affordance Learning via VAE Query Selection & Data Manipulation

Jordan Sinclair[1], Brian Reily[2] and Christopher Reardon[1]

***Abstract*— Terrain preference learning from trajectory queries allows complex reward structures to be obtained for robot navigation without the need for manual specification. However, traditional offline preference learning approaches suffer from ambiguous trajectory pairs stemming from inadequacy in the initial dataset, which causes longer learning times and may lead to less accurate results. Several approaches have been introduced to tackle this common problem including creating preference learning models robust to volatility in weights from ambiguous choices, enhancing the query selection process towards mitigating dubious trajectory choices, and modifying the original dataset with highly variant samples. Inspired by recent work in the application of deep learning towards improving query selection, this paper introduces a joint dataset and query optimization procedure utilizing variational autoencoders. Our efforts leverage both the encoder and decoder models to identify underrepresented terrain types and supplement the trajectory set with targeted samples created using the decoder. We jointly optimize a clustered latent space towards creating balanced clusters that can be used to obtain diverse trajectory pairs.**

I. Introduction

Robot navigation in nuanced environments with diverse features and terrain types necessitates well-structured costmaps that incorporate context-driven preferences. Traditionally weights were manually constructed for each unique terrain type by a human expert, but this quickly becomes intractable as the complexity of the environment and the task increases. Learning from Demonstration (LfD) [1], [2] is a technique that attempts to reduce the cognitive burden stemming from manual cost formation by inferring weights through expert provided demonstrations of expected behaviors. However, demonstrations can be costly and even dangerous to obtain, particularly in robotics scenarios.

Preference learning is an interactive method that learns rewards through user feedback over pairwise trajectory queries. Preference elicitation is an influential technique in the human-robot interaction (HRI) field [3] as it promotes information inference over simple choices rather than complex demonstrations, although each choice may elicit less knowledge than a demonstration [4]. There are many variations on the standard preference learning problem, including offline and online learning. Offline learning allows data collection to be done prior to user querying, which reduces latency and allows multiple sessions to be run on the same data for various contexts. However, offline preference learning is highly tied to the quality of the initial trajectory set. When data variance is very low, preference elicitation may yield inaccurate information and repeated queries, potentially causing significant shifts in the learned weights that do not align with user intents.

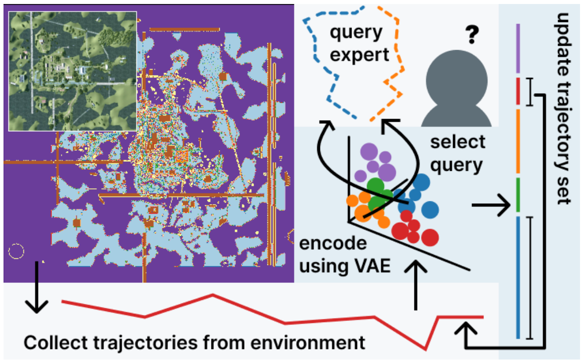

Fig. 1. System diagram outlining our approach: 1) trajectories T are collected from the environment, 2) Use trained encoder E to encode T into latent vectors, 4) Cluster these vectors into κ clusters representing κ terrain types, 5) Evaluate imbalance in cluster sizes and use the decoder D to generate new samples as needed that prioritize underrepresented terrain type γ_k corresponding to cluster C_k, 6) Pull trajectory pairs (τ_i, τ_j) from distinct latent clusters C_i, C_j, (7) display trajectories to user for preference elicitation.

In this work, we implement an interactive offline preference learning system for terrain weight assignment that leverages a generative model to enhance both the initial trajectory set and the query selection strategy, thereby increasing the overall diversity of the presented trajectories. Fig. 1 provides an overview of our proposed system. We validate our work with initial results via an integration with APReL [5], an extensive preference learning library directed at offline learning with a fixed trajectory set.

II. Background

Recent literature presents a substantial body of research focused on enhancing query distinction within low variance datasets. APReL [5] includes numerous query acquisition approaches, including mutual information, first introduced in [6]. This approach sought to strengthen query disparity through a focus on greedily selecting the trajectories with the highest expected information gain based on a cognitive model of the user. To further reduce the effects of indistinguishable queries, APReL [5] recently introduced batch querying [7], where a larger number of queries are presented to the user before updating the weights, making them more robust to feedback from unanswerable queries.

[8] proposed to use a dynamics model that learned transi-

*This work was supported by ARO W911NF-22-2-0238.

[1]Department of Computer Science, Ritchie School of Computer Science and Engineering, University of Denver, USA. {jordan.sinclair,christopher.reardon}@du.edu

[2], DEVCOM Army Research Laboratory, Adelphi, MD, USA brian.j.reily.civ@army.mil

tions in trajectory elements towards generating diverse experiences without the need for environmental interaction. This dataset was combined with a query determination process which created query batches, similar to [7], containing samples from past iterations to maintain variability. Rather than batches, [9] elicited preferences over queries from collected actions. This allowed unrepresentative policy changes to be discouraged with incoming feedback provided on a subset of collected data. [10] proposed to mitigate the effects of class imbalance in active learning by blending thee query selection methods leveraging two datasets, one unlabeled and the other labeled with preference information.

Other work in this area prioritizes the trajectory set itself, often employing a preliminary data manipulation stage to optimize variation for potential query selections. [11] used pre-training to obtain disparate experiences prior to preference querying. Those experiences were labeled with reward updates to continually provide diversified and information aligned data during preference elicitation. [12] employed advances in generative models for data augmentation by creating new trajectories stemming from initially unpreferred samples. This approach demonstrated improvements in query variance and introduced a more streamlined preference elicitation process by initially utilizing higher preference samples. Data modification prior to preference querying promotes a higher degree of information gain as trajectory pairs are more variegated, making queries easier to answer.

The application of generative models in preference learning can be utilized not only for data generation but also for enhancing query acquisition. [13] implemented a variational autoencoder (VAE) targeted towards minimizing repeated and uninformative queries in preference elicitation. The model was trained on trajectories, state-action pairs, formed from a reinforcement learning policy. The encoder of the VAE was used to form latent vectors from the initial trajectory set, which were then clustered. Trajectory pairs with high variability were obtained by sampling segments from distinct clusters. These queries were then ranked based on estimated information gains and presented to the user to elicit preferences.

III. Approach

We present a novel framework leveraging both the encoder and decoder of a variational autoencoder [12], [13] for improved query selection and data augmentation, respectively. Further, we employ our work towards terrain weight specification for context-driven robotic navigation scenarios. We demonstrate the efficacy of our combined approach through a comparison with the Mutual Information [6] acquisition function implemented in APReL [5]. Our overall approach can be seen in Fig. 1.

Consider a trajectory τ to be a sequence of L terrain types $\gamma_k \in \Gamma$ encoded using integer labels, and a trajectory set to contain N trajectories, $T = \tau_1, \tau_2, ..., \tau_N$. We utilized three distinct trajectory sets: 1) a training set T_t, 2) an initial trajectory set T_o for preference elicitation, and 3) an adjusted trajectory set T_D formed using the decoder D. Both components of a VAE, namely the decoder D and encoder E, are trained in unison using trajectory set T_t with standard reconstruction and KL-Divergence losses. We constructed the encoder and decoder models with Long Short Term Memory (LSTM) modules, along with linear layers to represent the outputs. This design allows for pattern recognition over long terrain sequences, important in generating realistic trajectories for preference learning. Training is completed prior to preference elicitation, so it doesn't impact user wait times and can be performed with a larger dataset T_t.

Once trained, the encoder E is used to transform trajectories $\tau_i \in T_o$ from the initial dataset into lower dimensional latent vectors. Specifically, the latent space of the VAE is represented with dimensions proportional to the number of distinct terrain types $\gamma_k \in \Gamma$. The encoded representations are then clustered using κ-means with $\kappa = |\Gamma|$, similar to the method presented in [13]. Different to this approach, however, our cluster assignment reflects a semantic correlation between terrain types $\gamma_k \in \Gamma$ and clusters C_k.

Rather than directly pulling from these clusters, as in [13], for preference queries, we propose a joint data augmentation and query set optimization stage. Specifically, we evaluate the balance between clusters to determine if any terrain types are not well represented. In that case, we use the decoder D to form a new trajectory set T_D by introducing new generated samples stemming from lacking clusters, thereby balancing the overall distribution of latent space clusters. Then queries are formed from distinct clusters C_i and C_j such that the trajectories contain a large amount of terrain types, γ_i and γ_j, respectively. This encourages diversity in the query suitable for preference elicitation.

We use APReL's [5] implementation of mutual information [6] for selecting suitable trajectory pairs to query the user with. This method is updated generally by constraining the sample space to consist only of trajectory pairs formed from distinct clusters, rather than all possible combinations.

IV. Preliminary Results

We validate the application of our method, joint VAE-assisted data enhancement and query selection, towards robot navigation via learnt cost assignment through a preliminary demonstration. We compare against the implementation of mutual information acquisition [6] provided by APReL [5]. We utilize a real-world 2D terrain map, see Fig. 1, consisting of five terrain types $\gamma_1, \gamma_2, ..., \gamma_5 \in \Gamma$, namely *water*, *sand*, *rock*, *trees*, and *sidewalk*.

Because preference learning is inherently biased, we opted to create a simulated user that mimics human preference feedback. Decisions are made simply based on which trajectory is most aligned with the ground truth rewards. In support of this, we introduce a preferential alignment metric that ranks the terrains in order of their coverage for a given trajectory and compares that with the ground truth reward (represented as a ranking of terrains as well). Alignment is then calculated as the number of terrains γ_k in the feature distribution of a trajectory τ that are out of order when compared to the ground truth representation.

The VAE was trained with $|T_t| = 320$ trajectory samples stemming from the same start and end positions. Each trajectory was formed by randomly selecting a direction to move (returning to the previous cell was excluded from the options at each stage) until either the goal position was reached or the trajectory had maximum length. We selected a maximum length of $L = |\tau_i| = 50$ cells, or 10 meters based on the map resolution, so each trajectory $\tau_1, \tau_2, ..., \tau_{320}$ had a length up to $L = 50$ cells. To maintain consistent terrain distribution for shorter trajectories, we padded the sequence by repeating earlier terrain types prior to training. This enhanced the realism of trajectories by allowing for variable lengths while maintaining consistent start and end points. The VAE was trained on the input set T_t for 5000 iterations.

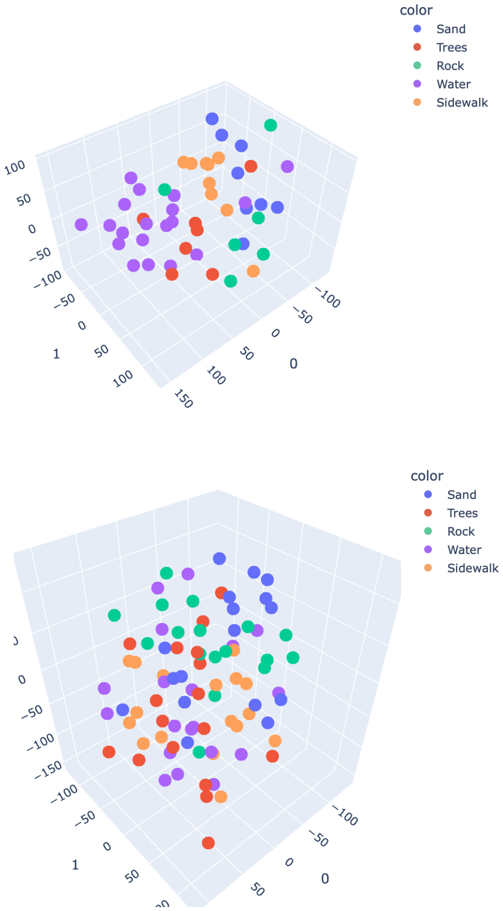

Fig. 2. (a) 3D T-SNE visualization of encoded trajectories from T_o. (b) 3D T-SNE visualization of 5D latent vectors, including additional decoded vectors from cluster balancing stage.

We compare our approach, as discussed previously, with a mutual information technique [6]. Fig. 3 shows our initial results with the training configuration discussed above and with an initial trajectory set with size $|T_o| = 50$. The data augmentation step added 37 trajectories to T_D, see Fig. 2. Preference learning with both query acquisition methods was conducted over 25 iterations. We consider a moving average over the five weights for each query as well as the reward alignment metric used in the simulated user. Consider the preferential ordering of the weights for each method on the last query. While neither approach obtained the exact ground truth ordering in the time allotted, our method was overall closer to an acceptable arrangement. Water was ranked close to the bottom in our results, as expected with its ground truth weight of -1, but near the higher end in the mutual information strategy. Similarly, sidewalk, the most positive reward in the ground truth configuration, was ranked second in our approach and last from mutual information. Fig. 3-c supports these semantic assessments as our approach showed significantly greater alignment at certain points during preference elicitation, indicating that queries with higher variance contributed to more information gain. Further, our proposed joint data enhancement and query selection optimization pipeline actuated in this experiment was able to reach ground truth terrain alignment, emphasizing a more detailed exploration of the user's mental model through more informative queries.

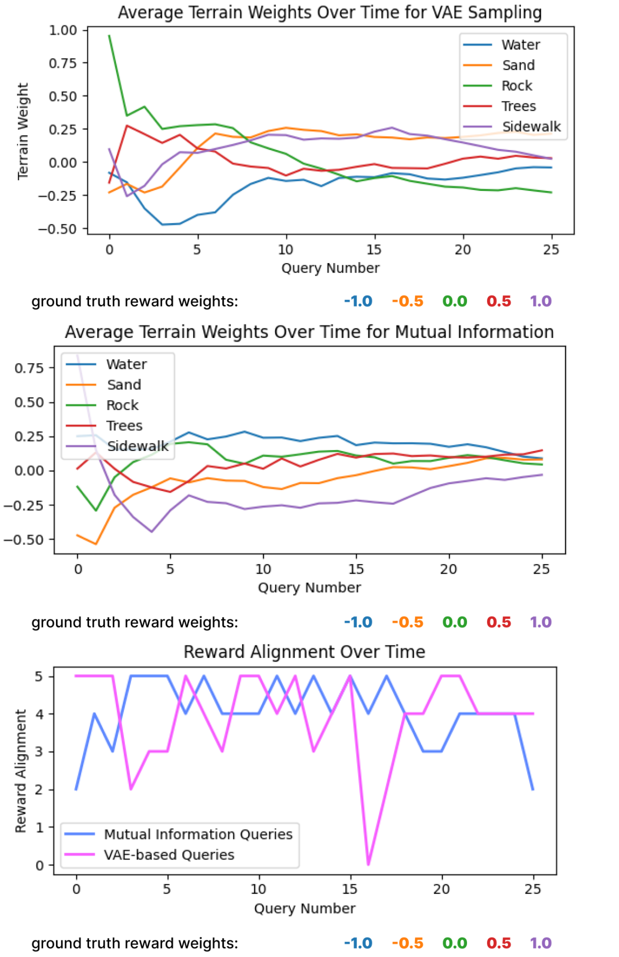

Fig. 3. (a) Moving average of terrain weights over VAE-enhanced querying. (b) Moving average of terrain weights over mutual information querying. (c) reward alignment over both querying strategies.

V. Conclusions & Future Work

In this paper we introduced a combined data enhancement and improved query acquisition approach utilizing both the encoder and decoder of a variational autoencoder model. We integrated our approach with existing state of the art preference learning approaches through an implementation of mutual information [6] in the APReL [5] library. Further, we applied both approaches towards learning terrain costs for robot navigation in a real environment, providing a useful alternative to manual specification and learning from demonstration [1], [2]. Our approach showed improvements over mutual information in few queries, even showing ground truth order alignment during the querying process. We hope to continue this direction of research by exploring different model configurations, including more advanced recurrent networks, to expand capabilities in generating realistic trajectories as well as to encode more complex structures. We also hope to integrate our approach with ROS (Robot Operating System) in order to use true path planning in more nuanced robotics environments with many terrain types.

References

[1] D. Silver, J. A. Bagnell, and A. Stentz, "Learning from demonstration for autonomous navigation in complex unstructured terrain," *International Journal of Robotics Research (IJRR)*, vol. 29, no. 12, pp. 1565–1592, 2010.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. Volume 3, 2020, pp. 297–330, 2020.

[3] Y. Abdelkareem, S. Shehata, and F. Karray, "Advances in preference-based reinforcement learning: A review," in *International Conference on Systems, Man, and Cybernetics (SMC)*, 2022.

[4] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," in *International Conference on Neural Information Processing Systems*, 2018.

[5] E. Bıyık, A. Talati, and D. Sadigh, "Aprel: A library for active preference-based reward learning algorithms," in *International Conference on Human-Robot Interaction (HRI)*, 2022.

[6] E. B, M. Palan, N. C. Landolfi, D. P. Losey, and D. Sadigh, "Asking easy questions: A user-friendly approach to active reward learning," in *Proceedings of the Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 1177–1190.

[7] E. Bıyık, N. Anari, and D. Sadigh, "Batch active learning of reward functions from human preferences," *Journal of Human-Robot Interaction*, February 2024.

[8] Y. Liu, G. Datta, E. Novoseller, and D. S. Brown, "Efficient preference-based reinforcement learning using learned dynamics models," in *International Conference on Robotics and Automation (ICRA)*, 2023.

[9] Q. Yang, S. Wang, M. G. Lin, S. Song, and G. Huang, "Boosting offline reinforcement learning with action preference query," in *International Conference on Machine Learning*. JMLR, 2023.

[10] G. Kim and C. D. Yoo, "Blending query strategy of active learning for imbalanced data," *IEEE Access*, vol. 10, pp. 79 526–79 542, 2022.

[11] K. Lee, L. Smith, and P. Abbeel, "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," in *International Conference on Machine Learning*, 2021.

[12] Z. Zhang, Y. Sun, J. Ye, T.-S. Liu, J. Zhang, and Y. Yu, "Flow to better: Offline preference-based reinforcement learning via preferred trajectory generation," in *International Conference on Learning Representations*, 2024.

[13] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Variquery: Vae segment-based active learning for query selection in preference-based reinforcement learning," in *International Conference on Intelligent Robots and Systems (IROS)*, 2023.