Learning to Discern: Imitating Heterogeneous Human Demonstrations with Preference and Representation Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract: Practical Imitation Learning (IL) systems rely on large human demonstration datasets for successful policy learning. However, challenges lie in maintaining the quality of collected data and addressing the suboptimal nature of some demonstrations, which can compromise the overall dataset quality and hence the learning outcome. Furthermore, the intrinsic heterogeneity in human behavior can produce equally successful but disparate demonstrations, further exacerbating the challenge of discerning demonstration quality. To address these challenges, this paper introduces Learning to Discern (L2D), an imitation learning framework for learning from demonstrations with diverse quality and style. Given a small batch of demonstrations with sparse quality labels, we learn a latent representation for temporally-embeded trajectory segments. Preference learning in this latent space trains a quality evaluator that generalizes to new demonstrators exhibiting different styles. Empirically, we show that L2D can effectively assess and learn from varying demonstrations, thereby leading to improved policy performance across a range of tasks in both simulation and on a physical robot.

Keywords: Imitation Learning, Preference Learning, Manipulation

1 Introduction

Imitation Learning (IL) allows robots to learn complex manipulation skills from offline demonstration datasets [1, 2, 3, 4, 5]. However, the quality of the demonstrations used for IL will significantly influence the effectiveness of the learned policies [2, 6, 7]. Practical imitation learning systems must amass demonstrations from a broad spectrum of human demonstrators [8, 9], ranging from novices to experts in a given task [10, 11]. Moreover, even experts may complete a task differently, resulting in disparate demonstrations of equal quality [12, 13, 14]. Therefore, an effective imitation learning algorithm must learn from demonstrations of varying quality, each further diversified by the unique skillset of the individual demonstrators.

However, mainstream imitation learning algorithms often make the simplifying assumption that all demonstrations are uniformly ideal [11]. As a result, policies trained with these algorithms may unknowingly learn from suboptimal or even contradictory supervisions [11, 14, 15]. Recent works have attempted to estimate expertise by developing unsupervised algorithms to quantify action distribution and variance [6] or actively soliciting new demonstrations that match with a known expert policy [16]. While these methods are general in principle, two critical limitations remain. First, estimating demonstrator expertise without any external quality label is an inherently ill-posed problem. As noted above, demonstrated behaviors can vary drastically across demonstrators and even across different trials from the same demonstrator. Neither action variance nor similarity to an expert is sufficient to capture expertise in such settings. Second, these work only consider state-level features, while behaviors are better represented in temporal sequences.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

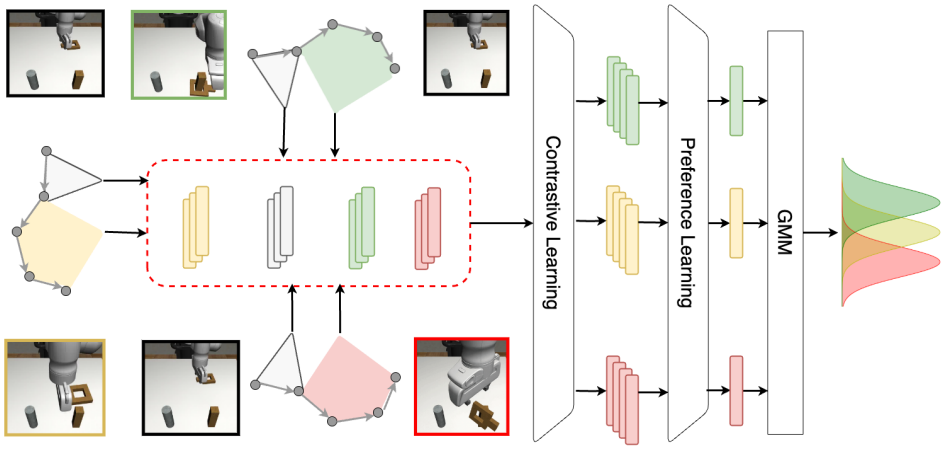

Figure 1: **L2D:** Our framework proceeds in three primary stages during training. First, we augment trajectory segments with temporal embeddings and employ contrastive learning to map these segments to a latent space. Next, we use preference learning in this latent space to train a quality critic on sparse preference labels. Finally, we train a Gaussian Mixture Model (GMM) on the critic's outputs where the different modes represent demonstrator quality.

To this end, we present Learning to Discern (L2D), a method to learn from heterogeneous demonstrations. L2D can efficiently estimate expertise from limited demonstration quality labels based on sequence-level latent features. By incorporating temporal embeddings into trajectory segments, we are able to learn a more meaningful latent representation for long-horizon manipulation tasks. Preference learning in this latent space trains a quality evaluator that generalizes across diverse styles of task completion while only requiring a subset of coarsely-ranked data. We demonstrate that our method can not only discern the quality of in-domain data, but also identify high-quality demonstrations from entirely new demonstrators with modes of expertise unseen in training. We show this empirically in both simulated and real-world robot settings. With L2D, it becomes feasible to filter high-quality demonstrations from the vast and diverse datasets that are needed for data-driven Imitation Learning.

2 Related Work

Our work intersects several domains, including imitation learning with suboptimal data, preference learning from human demonstrations, and representation learning for data retrieval.

Imitation Learning with Suboptimal Data. Imitation Learning (IL) [17, 18, 19, 20, 21, 22] from sub-optimal data is an active area of research in robotics. Various works show that near-optimal policies can be trained if demonstrations are ranked based on their quality [4, 11]. However, manual annotation of dense reward labels is costly and unscalable. CAIL [23] extrapolates demonstration quality from a subset of labeled data D', but still requires dense ranking of D'. Other works estimate demonstrator quality in an unsupervised manner (ILEED [6]) or compare demonstrations to an expert policy (ELICIT [16]). Such unsupervised algorithms do not generalize to equal quality demonstrations from different modes of task completion. This motivates the need for an IL algorithm that learns a generalizable quality estimator from a subset of coarsely-ranked data.

Preference Learning from Human Demonstrations. Preference learning is a form of supervised learning that learns rankings from pairwise comparisons [24, 25, 26, 27]. When applied to human demonstration data, preference learning can effectively train a reward function based on rankings of trajectory segments [28]. Various frameworks utilize this reward function to train policies through inverse reinforcement learning [29, 30]. To limit manual annotations, some methods synthetically generating lower quality data through noise injection [7]. These methods are hampered by the fact

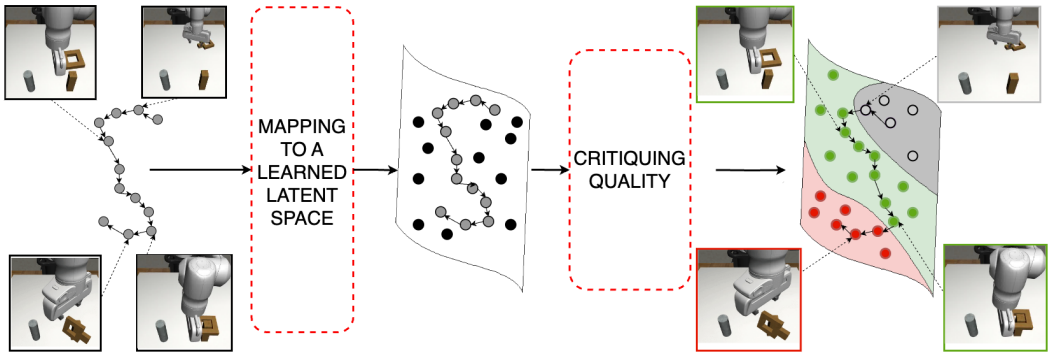

Figure 2: **Filtering Unseen Demonstrations:** When faced with unseen demonstrations, L2D partitions the trajectory into segments and augments each with it's chronological ordering in the sequence. The segments are mapped to the latent space learned during training and ranked by the quality critic. After calculating the mean and variance of ranks in a full trajectory, the trained GMM is employed to predict a preference label for the unseen demonstration.

that noise level does not directly translate to preference ranking of trajectories [31]. Our method, L2D, ranks offline demonstrations using preference learning in a learned latent space. This allows us to generalize across diverse styles and modes of task completion with sparse preference labels.

Representation Learning for Data Retrieval. Representation learning, especially with a focus on robotics, has been widely studied. This area typically explores learning from visual inputs [32], data augmentation [33], goal-aware prediction [34], and domain-specific information [35]. Recent work such as that by Du et al. [36] enhances the performance of robot manipulation tasks by first mapping offline data to a latent space using contrastive learning and then retrieving offline data using that latent space with task-specific data as queries. Yet, these works do not account for the suboptimality of given demonstrations in task-specific data and offline data.

3 Problem Setting

We study the imitation learning problem in environments modeled as Markov Decision Processes (MDP), where $\mathcal{S}$ and $\mathcal{A}$ denote the state space and the action space, respectively. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the state transition function. The environment emits binary reward upon task completion. The learning algorithm does not have access to this signal.

We assume access to a small set of demonstration $D_{\text{known}} = \tau_{i}{}_{i=1}^{N}$, where a trajectory τ_i is a sequence of length-L transitions $(s_0, a_0, s_1, a_1..., s_{L-1}, a_{L-1})$. Each trajectory in D_{known} is associated with a quality label $l \in \mathcal{L}$, where $\mathcal{L}$ is the set of possible labels indicating the quality of a trajectory. Based on these quality labels, we divide D_{known} into sub-datasets of similar quality levels, i.e., $D_A, D_B, D_C, \ldots$. Each sub-dataset, such as D_A, comprises a collection of trajectory segments $\tau_A^1, \tau_A^2, \ldots, \tau_A^T$. If dataset A is considered to be of superior quality compared to B, we denote it as $A \succ B$. Likewise, we assume $\tau_A \succ \tau_B$, provided $\tau_A \in A$ and $\tau_B \in B$.

In this paper, we develop a framework for imitation learning that achieves state of the art performance on sub-optimal data by estimating the quality of new demonstrations. More specifically, our aim is to learn a representation that captures the multimodality from a large dataset of unknown quality D_{unknown} and can critique the quality of new demonstrations in the same.

4 Learning to Discern

Imitation Learning (IL) with suboptimal data often leads to inferior task performance. Evaluating the quality of a demonstrator without supervision is challenging, and manually assigning quality to demonstrations is impractical. The core component of our method is a preference network Q that learns to evaluate the quality of a demonstration τ by estimating its ranking label l. As mentioned above, the key challenges of learning a generalizable demonstration quality estimator are that (1) it is difficult to capture behavior-level feature by assessing individual state-action pairs, especially for long-horizon manipulation tasks and (2) the demonstrated behaviors are diverse even among demonstrators of similar expertise. In this section, we describe each challenge in more details and introduce the components in our method L2D that address the corresponding challenges.

Extracting Behavior Feature through Temporal Contrastive Learning

As we mentioned earlier, estimating the quality of a demonstration by merely observing variance in states is a poorly-posed problem. Instead, we propose learning a temporal latent space for quality critique. We utilize a neural network encoder $E : \mathbb{R}^{L_1 \times \text{Obs.Dim}} \rightarrow \mathbb{R}^{\text{Latent.Dim}}$ that maps trajectory segments $\{\sigma_j = \{s_t\}_{j,t=i}^{t=i+L_1}\}_{j=1}^{i=\text{num_segments}}$ into a d-dimensional latent space. Sampling triplet pairs for this learning poses a challenge: naive sampling based on quality labels may degrade representation quality. Specifically, long-horizon tasks like Square contain bottleneck states indifferent to demonstration quality, leading to task-specific regions that don't majorly impact overall quality. Thus, we devise a custom sampler to get quality influencing trajectory segments along with quality independent trajectory segments. We use these sampled segments to create triplets of an anchor trajectory segment σ_a, a positive trajectory segment σ_p and a negative trajectory segment σ_n which are used in different contrastive learning strategies:

1. **Strategy 1: Arbitrary Sampling from Distinct Quality Sets** We sample σ_a and σ_p from different trajectories from a demonstration set A. We then sample the negative trajectory segment σ_N from a different demonstration quality set B.
2. **Strategy 2: Specific Segment Sampling from Arbitrary Quality Sets**We sample the region specific segments like initial segments $\sigma_{a,pos}$ and $\sigma_{p,pos}$ from two different trajectories τ^1 and τ^2, respectively. The negative trajectory segment σ_n can not be the from same region of any trajectory. For example, in case of square task with all complete demonstrations, we leveraging that all demonstrations end with the the robot placing the square in the cylinder. Consequently, we sample the final segments $\sigma_{a,final}$ and $\sigma_{p,final}$ from two different trajectories τ^1 and τ^2, respectively. The negative trajectory segment σ_n can not be the final segment from any trajectory. These trajectories segments belong to any demonstration set as the segmented regions are not quality influencing. Please see the Table 2 for an ablation.

Position Encoding for Capturing Non-cyclic Behavior

Preference learning effectively identifies demonstration quality in repetitive tasks like Hopper [37], but this success is not mirrored in long-horizon manipulation tasks. This discrepancy, we believe, is due to the non-repetitive nature of the latter tasks, where similar segments at different trajectory points may imply different demonstration qualities. To accommodate this variability, we introduce data augmentations like position encoding for each demonstration state with respect to the entire trajectory. We find that this provies the necessary context to the latent space for comprehensive task understanding. Let our original observation at any time-step t be denoted as $o_t \in \mathbb{R}^{obs.dim}$. We introduce position encoding, represented as a real number p_t indicating the normalized time-step (i.e., t/T, where T is the total number of time-steps in the demonstration). We then formulate the augmented observation o'_t which becomes a vector in $\mathbb{R}^{obs.dim+1}$, defined as $o'_t = [o_t, p_t]$.

Training Quality Critic

After training the encoder E to learn a latent space that is cognizant of quality sensitive regions of the demonstrations, we train a quality critic network $Q : \mathbb{R}^d \rightarrow \mathcal{L}$ that takes a trajectory segment

trajectory embedding as input and predicts its quality label. We take pairs of randomly sampled trajectory segments σ_A and σ_B from different datasets D_A and D_B. Assuming $D_A \succ D_B$, then $\sigma_A \succ \sigma_B$ or vice versa. Similar to [29], we use 10% label noise during training Q.

Respresenting Suboptimality as a Gausian Mixture Model

Demonstrations of varying quality may exhibit mixtures of behaviors in different regions. By being receptive to individual demonstration regions, our approach allows us to estimate the quality of these mixed-quality demonstrations, as the overall quality of a demonstration can be viewed as a majority vote of its constituent parts. This concept can be naturally represented by a Gaussian distribution, where the mean of the Gaussian represents the average quality of segments in a demonstration. After learning the latent space encoder E and quality critic Q, we train a Gaussian Mixture Model (GMM) G. We sample a sufficient number of segments ($\lceil \frac{L}{L_1} \times k \rceil$) from each demonstration in D_{known} and pass each segment through encoder E and quality critic Q to obtain a scalar value. These values represent the quality of different regions within the demonstration. We then map these scalar values into sets that correspond to demonstrations of different qualities (i.e., D_A, D_B, D_C, etc) and train the GMM where different gaussians correspond to different qualities.

Estimating Quality of Unseen Demonstrations Given a new unseen demonstration τ, we generate a set of scalar values using $Q(E(\text{sampler}(\tau)))$, with each value estimating the quality of a specific region within the demonstration. We then estimate the probability of each value with respect to the trained GMM and assign each scalar to the set with the maximum probability. This gives a comprehensive quality estimate for different regions of the demonstration. Finally, we use a heuristic to determine whether the demonstration should be used for training the policy. For example, by counting the number of segments assigned to the highest quality set, we can select demonstrations if the majority exhibit good behavior.

5 Experimental Evaluation

We empirically validate the whether L2D can learn effective representations that enables accurate discrimination of high-quality demonstrations from the pool of available demonstrations, which vary in quality. We show that distinct components of multi-compenent architecture of L2D significantly contribute to its effectiveness with an ablation study. We demonstrate that L2D can discern high-quality demonstrations from both seen and unseen demonstrators and compare against baselines. Finally, we collect human demonstrations for real and simulated manipulation tasks to test the method's applicability in more realistic scenarios.

Experiment setup. We use the Robomimic dataset for all experiments, which includes a range of tasks performed by different human operators. We specifically focus on the Square Nut Assembly task where Robomimic provides us with 300 demonstrations. These demonstrations are categorized into 'good', 'ok', or 'bad' based on the execution quality, and are collected from 6 different demonstrators. Once we have filtered demonstrations, we employ the Behaviour Cloning (BC-RNN) algorithm for training and report the success rate for different methods.

Baselines. We benchmark the performance of our proposed method, L2D, against two state-of-the-art method for learning from mixed-quality demonstrations. ILEED [6] employs an unsupervised expertise estimation approach to identify good demonstrations. ELICIT [16] actively filters new demonstrations based on whether the state-action pairs match the prediction of an expert policy. The naive baseline is a simple uniform sampling strategy from D_{unknown}. The oracle baseline stands at the opposite end of the spectrum, representing an ideal scenario where the highest-quality demonstrations can be perfectly identified from D_{unknown}.

5.1 Component Ablation

In our ablation study, we dissect L2D to evaluate the individual contribution of each component.By comparing the Wasserstein distances between quality subset distributions within D_{unknown}, we quan-

tify the quality of the representation learned. Higher total Wasserstein distances signify more effective representations. Our findings, as illustrated in Table 3 and corresponding distribution visualizations, reveal that L2D consistently outperforms the basic preference learning approach by achieving higher Wasserstein distances. Adding the S1 Contrastive component and positional encoding enhances the quality of the learned latent space, confirming the necessity of contextual understanding for long horizon non-cyclic robot manipulation tasks. The final addition of the S2 Contrastive component with data augmentations like time warping underlines the benefits of using well-crafted positive and negative pairs for contrastive learning. These pairs, even from demonstrations of differing quality, lead to better representations, highlighting the importance of understanding that some regions of demonstrations do not influence the quality. This lends support to the idea of leveraging commonalities across all demonstrations to regularize our learning process, thereby avoiding overfitting to specific demonstrator biases.

Method	**G vs. O** ↑	**G vs. B**↑	**O vs. B**↑	**Total Distance**↑
Preference	0.130	0.143	0.147	0.420
+S1 Contrastive	0.337	0.213	0.414	0.964
+ Position Encoding	0.535	0.649	0.589	1.773
+ S2 Contrastive+DataAug	0.604	0.848	0.417	**1.869**

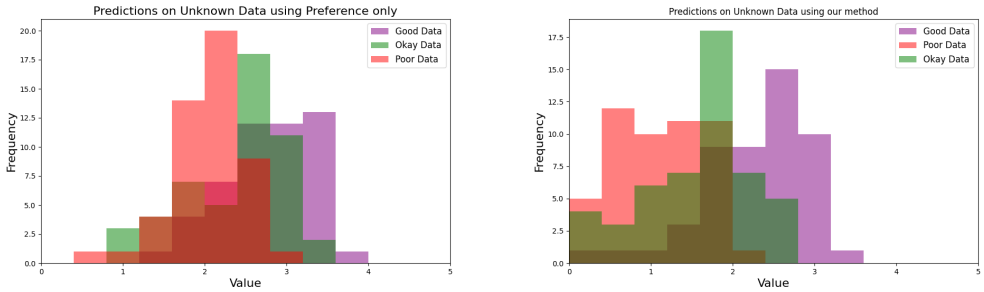

Figure 3: **Method vs. Wasserstein Distances and Distribution Visualization:** We measure the separability of estimated demonstration quality from new demonstrators of varying levels of quality (good, okay, bad) on the Robomimic Square task.

5.2 Main Results

L2D can identify unseen high-quality demonstrations from familiar demonstrators. We assess the efficacy of L2D in estimating the quality of demonstrations provided by demonstrators seen in training. For the purpose of this experiment, we partition the 300 publicly available suboptimal demonstrations from the Robomimic Square Nut Assembly task into two sets: D_{known} and D_{unknown}, each comprising 150 demonstrations. These sets represent similar variations of suboptimal demonstrations, sourced from identical demonstrators. Our results indicate that L2D successfully identifies 43 out of the top 50 high-quality demonstrations from the pool. We find that our method outperforms [6] reinforces the idea that using preferences over an unsupervised approach leads to better representations. Finally, we see that our method achieves oracle level SR when training the policy on filtered data.

L2D can identify high-quality demonstrations from unseen demonstrators. In practical situation, a policy may be trained on data provided by previously unseen demonstrators. This is challenging because each demonstrator may possess a unique style of execution and a diverse range of demonstration quality. To evaluate L2D in this setting, we again split the 300 demonstrations from the Robomimic dataset into two sets, but this time, the demonstrations in each set originate from different demonstrators. We then use our method to filter the top 50 demonstrations from all methods for training the imitation learning policy.

Results in Figure 5 show that while both contrastive learning and preference learning methods perform better than naive sampling, our proposed method L2D significantly outperforms both. This

Method	Success Rate
Naive	0.44
ILEED	0.54
Our Method	0.66
Oracle	0.66

Method	Good	Okay	Bad
Naive	68	16	16
Our Method	93	4	3
Oracle	100	0	0

Figure 4: **Identifying Unseen High-Quality Demonstrations from Familiar Demonstrators:** We use different methods to select high quality demonstrations from the unknown dataset and report the quality of the selected demos on the right, along the the success rate of imitation learning policy when train on selected demos along with best known demonstrations. We find that our method's performance is on par with the oracle.

implies that L2D is more adaptable to variation in demonstration quality and demonstrator style, as it effectively selects a majority of the high-quality demonstrations. In comparison with the ELICIT method [16], which learns only from high-quality data, L2D demonstrates superior performance. Notably, the success rate achieved by our method is only marginally inferior to that of the oracle, further supporting our claim that L2D can effectively discern high-quality demonstrations from unfamiliar demonstrators.

Method	Success Rate
Naive	0.20
Contrastive	0.36
Preference	0.38
Elicit	0.38
Our Method	**0.44**
Oracle	0.46

Method	Good	Okay	Bad
Naive	18/50	16/50	16/50
Contrastive	23/50	20/50	7/50
Preference	27/50	20/50	3/50
Elicit	23/50	20/50	7/50
Our Method	**39**/50	**8**/50	**3**/50
Oracle	50/50	0/50	0/50

Figure 5: **Identifying Unseen High-Quality Demonstrations from New Demonstrators:** We use different methods to select high quality demonstrations from the unknown dataset and report the quality of the selected demonstrations in the table on the right.

5.3 Learning from Real-World Demonstrations

In this subsection, we evaluate the aptitude of L2D in estimating the quality of demonstrations outside of the standard benchmark, and using these demonstrations to improve the success rate (SR) of task performance. This evaluation is geared towards validating the effectiveness of our method in more realistic scenarios in comparison to controlled simulated environments.

We gathered new human demonstrations for tasks such as Square Nut Assembly, Lifting, and Stacking in both simulated and real-world environments, utilizing a diverse group of operators. These operators performed independently, without leveraging knowledge from others' demonstrations. We then trained our method on a subset of these demonstrations and utilized it to identify high-quality demonstrations from the unseen dataset. Note that in the real world, both the preference model and the policy are trained end-to-end with image-based demonstration data. Please refer Sec. 11 in Appendix for more training details. The effectiveness of our method is evaluated based on the improvement in SR after the integration of these newly identified demonstrations. Our experimental results, depicted in the table below, suggest that L2D substantially improves the performance of imitation learning policy in simulation and real-world scenarios. We visualize the quality prediction for two stacking demonstrations and show that L2D can correctly identify high-quality demonstrations based on sparse labels.

6 Limitations

In this work, we have recognized and addressed the challenges associated with discerning human demonstration datasets, which typically exhibit a wide variance in terms of quality and styles of

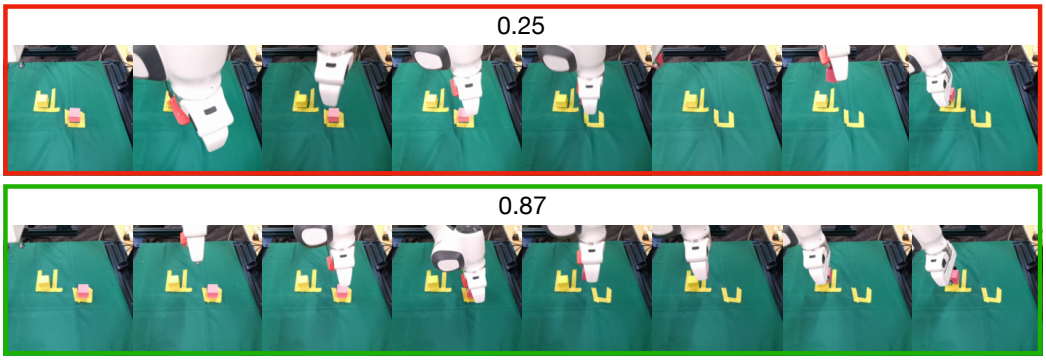

Figure 6: **Demo Quality Estimation.** We show the predicted quality for real-world demonstrations of the Stack task. The trajectory segments that with less optimal behaviors (e.g., jittering or waving) will be assigned with lower scores (marked with red bounding boxes).

Environment	Task	Method	High-Quality Demo Selection (%)	SR (%)
Simulated	Square	Naive Sampling	37	10
Simulated	Square	Our Method	**95**	**32**
Real	Stack	Naive Sampling	33	50
Real	Stack	Our Method	**97**	**80**
Real	Lift	Naive Sampling	33	30
Real	Lift	Our Method	**45**	**53**

Table 1: Performance comparison of naive sampling and our method in terms of high-quality demonstration selection from an unknown dataset and SR increase for two tasks (Square Nut Assembly and Pick and Place) in a simulated environment and one task (Lifting) in a real environment.

demonstrations. While our approach effectively mitigates demonstrator bias and learns to generate generalized representations, it still contains certain limitations.

Our method's reliance on filtering out entire demonstrations based on their assessed quality. This approach may be overly simplistic, as it assumes that a low-quality demonstration does not contain any high-quality segments. Demonstrations often encompass a diverse array of tasks and behaviors, and even a low-quality demonstration may contain sections that could provide useful training data. Additionally, our method assumes a clear distinction between high and low-quality demonstrations. However, the line separating high-quality from low-quality demonstrations is often blurred, depending on the complexity of the task and the specificity of the demonstrator's style. Inculcating this into L2D can lead to even better learned representations.

7 Conclusion

We presented Learning to Discern (L2D), a novel method for efficiently estimating expertise from heterogeneous demonstration data in IL. L2D employs sequence-level latent features and temporal embeddings to learn a robust latent representation of complex tasks. With a quality evaluator trained in this latent space, our method can generalize across varied task completion styles and identify high-quality demonstrations, even from new, unseen modes of expertise. This study demonstrates the potential of L2D to filter high-quality demonstrations from vast and diverse datasets, marking a significant advancement in the field of data-centric robotics.

References

[1] S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643, 2020.

[2] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. arXiv preprint arXiv:2108.03298, 2021.

[3] S. Emmons, B. Eysenbach, I. Kostrikov, and S. Levine. Rvs: What is essential for offline rl via supervised learning? arXiv preprint arXiv:2112.10751, 2021.

[4] A. Kumar, J. Hong, A. Singh, and S. Levine. When should we prefer offline reinforcement learning over behavioral cloning? arXiv preprint arXiv:2204.05618, 2022.

[5] D. Ghosh, A. Ajay, P. Agrawal, and S. Levine. Offline rl policies should be trained to be adaptive. In International Conference on Machine Learning, pages 7513–7530. PMLR, 2022.

[6] M. Beliaev, A. Shih, S. Ermon, D. Sadigh, and R. Pedarsani. Imitation learning by estimating expertise of demonstrators. In International Conference on Machine Learning, pages 1732–1748. PMLR, 2022.

[7] D. S. Brown, W. Goo, and S. Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In Proceedings of the 3rd Conference on Robot Learning, 2019.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022.

[9] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. arXiv preprint arXiv:1909.12200, 2019.

[10] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In Conference on Robot Learning, pages 879–893. PMLR, 2018.

[11] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In Conference on Robot Learning (CoRL), 2021.

[12] A. Mandlekar*, **D. Xu***, R. Martín-Martín, S. Savarese, and L. Fei-Fei. Learning to generalize across long-horizon tasks from human demonstrations. Robotics: Science and Systems, 2020.

[13] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto. Behavior transformers: Cloning k modes with one stone. Advances in neural information processing systems, 35:22955–22968, 2022.

[14] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. arXiv preprint arXiv:2303.04137, 2023.

[15] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson. Implicit behavioral cloning. In Conference on Robot Learning, pages 158–168. PMLR, 2022.

[16] K. Gandhi, S. Karamcheti, M. Liao, and D. Sadigh. Eliciting compatible demonstrations for multi-human imitation learning. In K. Liu, D. Kulic, and J. Ichnowski, editors, Proceedings of The 6th Conference on Robot Learning, volume 205 of Proceedings of Machine Learning Research, pages 1981–1991. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/gandhi23a.html.

[17] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. Robotics and autonomous systems, 57(5):469–483, 2009.

[18] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. Neural computation, 3(1):88–97, 1991.

[19] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[20] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In International conference on machine learning, pages 49–58. PMLR, 2016.

[21] J. Ho and S. Ermon. Generative adversarial imitation learning. Advances in neural information processing systems, 29, 2016.

[22] Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp. Goal-conditioned imitation learning. Advances in neural information processing systems, 32, 2019.

[23] S. Zhang, Z. Cao, D. Sadigh, and Y. Sui. Confidence-aware imitation learning from demonstrations with varying optimality. Advances in Neural Information Processing Systems, 34: 12340–12350, 2021.

[24] T. Joachims. Optimizing search engines using clickthrough data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 133–142, 2002.

[25] J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 14, pages 145–156. Springer, 2003.

[26] W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In Proceedings of the 22nd international conference on Machine learning, pages 137–144, 2005.

[27] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.

[28] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

[29] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 783–792, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/brown19a.html.

[30] C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, et al. A survey of preference-based reinforcement learning methods. Journal of Machine Learning Research, 18(136):1–46, 2017.

[31] L. Chen, R. Paleja, and M. Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In Conference on robot learning, pages 1262–1277. PMLR, 2021.

[32] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. arXiv preprint arXiv:2203.12601, 2022.

[33] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. Advances in neural information processing systems, 33:19884–19895, 2020.

[34] S. Nair, S. Savarese, and C. Finn. Goal-aware prediction: Learning to model what matters. In International Conference on Machine Learning, pages 7207–7219. PMLR, 2020.

[35] R. Jonschkowski and O. Brock. Learning state representations with robotic priors. Autonomous Robots, 39:407–428, 2015.

[36] M. Du, S. Nair, D. Sadigh, and C. Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets. arXiv preprint arXiv:2304.08742, 2023.

[37] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. arxiv. arXiv preprint arXiv:1606.01540, 10, 2016.

[38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

8 Supplementary Ablation Studies: Trajectory Similarity and Time Warping

In addition to the ablation studies discussed in the main body of the paper, we also investigate the effect of different architectural components and design choices in our proposed method. We focus specifically on the roles of initial and final trajectory segment similarity and the usage of location sensitive time warping.

To conduct this ablation study, we again select the square task from the robomimic dataset, utilizing its publicly available suboptimal human demonstrations. We train our method using demonstrations from three demonstrators of varying quality and test its generalizability using demonstrations from the remaining three different demonstrators. Both D_{known} and D_{unknown} contain demonstrations of similar suboptimality but sourced from different demonstrators.

In these experiments, we measure the Wasserstein distances between trajectories when employing an LSTM+Self Attention-based architecture for contrastive learning, utilizing a CLS token followed by a simple MLP encoder for preference learning. We systematically examine the increase in separability of demonstration quality from new demonstrators when incorporating our method's design choices, such as initial trajectory segment similarity, final trajectory segment similarity, and location sensitive time warping.

S2.Contrastive	**S3.Contrastive**	**Data Aug**	**G vs O**	**G vs B**	**O vs B**	**Sum**(↑)
No	No	No	0.12	0.24	0.36	0.72
Yes	No	No	0.19	0.28	0.4	0.87
Yes	No	Yes	0.16	0.4	0.36	0.92
Yes	Yes	Yes	0.28	0.48	0.48	**1.24**

Table 2: **Impact of Trajectory Segment Similarity and Location Sensitive Time Warping:** The ablation study demonstrates the value added by incorporating initial and final trajectory segment similarity and location sensitive time warping into our method. The table highlights that each of these components enhances the separability of demonstration quality from new demonstrators, leading to superior representations and increased total Wasserstein distance.

Table 2 presents the results of our supplementary ablation study, focusing on the impact of trajectory similarity and location sensitive time warping. As each of these elements is integrated into our method, we observe an increase in total Wasserstein distance, indicating an improvement in the separability of demonstration quality. The final configuration, which includes initial and final trajectory similarity and location sensitive time warping, achieves the highest total Wasserstein distance, reinforcing the effectiveness of our method's design choices in improving the quality of the learned representations.

9 Real Robot Demonstrations

Figure 7: **Real Robot Execution.** We show the rollout results of the policies trained with the demonstrations selected by our method and naive sampling for the Stack task.

We present visualizations of the real-world execution outcomes achieved by policies learned using demonstrations identified through our approach and naive sampling for the Stack task. Our findings show that policies learned from higher quality demonstrations exhibit robust and efficient task completion, achieving an 80% success rate over 10 trials. On the other hand, policies learned from demonstrations selected through naive sampling occasionally display peculiar behaviors, such as

waving around the target position or dropping the grasped cube at unsafe heights. Consequently, these behaviors contribute to a lower success rate of 50%.

10 L2D Hyperparams

Hyperparameter	Default
discriminator_learning_rate	0.0001
discriminator_num_segments	20000
discriminator_segment_length	48
discriminator_batch_size	128
discriminator_training_steps	500000
initial_trajectory_segment_len	12
final_trajectory_segment_len	6
use_pos_encoding	True
discriminator_embedding_size	12
Architecture	[LSTM,2 MLPs,Flatten(), 2 MLPs] [2MLPs]

Table 3: Hyperparameters used for training L2D to perform Robomimic based ablations.

11 Policy Training Details

We choose the default BC-RNN model setting from Robomimic benchmark [11] for learning policies. For simulation tasks the agents only use low-dimensional observation (e.g., robot proprioception, object poses), the epochs are set to 2000 with batch size 100, we test the agents by running rollout with a maximal horizon of 500 in every 50 epochs. For real-world tasks, we also include the RGB color observations from the three cameras (two at the side, one at the robot's wrist) and the epochs are set to 500. All networks are trained using Adam optimizers [38].

Hyperparameter	Default
Batch Size	16
Learning rate (LR)	1e-4
Num Epoch	1000
Train Seq Length	10
MLP Dims	[400, 400]
Image Encoder - Wrist View	ResNet-18
Image Feature Dim	64
GMM Num Modes	5
GMM Min Std	0.01
GMM Std Activation	Softplus

Table 4: Hyperparameters for L2D