
Representation Learning in Deep RL via Discrete Information Bottleneck

Riashat Islam

Mila, McGill University
Microsoft Research Montreal

Hongyu Zang

Beijing Institute of Technology

Manan Tomar

AMII, University of Alberta
Microsoft Research Montreal

Aniket Didolkar

Mila, University of Montreal

Md Mofijul Islam

University of Virginia

Samin Yeasar Arnob

Mila, McGill University

Tariq Iqbal

University of Virginia

Xin Li

Beijing Institute of Technology

Anirudh Goyal

Google DeepMind

Nicolas Heess

Google DeepMind

Alex Lamb

Microsoft Research NYC

Abstract

Several self-supervised representation learning methods have been proposed for reinforcement learning (RL) with rich observations. For real world applications of RL, recovering underlying latent states is crucial, particularly when sensory inputs contain irrelevant and exogenous information. In this work, we study how information bottlenecks can be used to construct latent states efficiently in the presence of task irrelevant information. We propose architectures that utilize variational and discrete information bottlenecks, coined as REPDIB, to learn structured factorized representations. Exploiting the expressiveness bought by factorized representations, we introduce a simple, yet effective, bottleneck that can be integrated with any existing self supervised objective for RL. We demonstrate this across several online and offline RL benchmarks, along with a real robot arm task, where we find that compressed representations with REPDIB can lead to strong performance improvements, as the learnt bottlenecks help predict only the relevant state, while ignoring irrelevant information.

1 Introduction

In the most general reinforcement learning (RL) setting, an agent is tasked with discovering a policy that achieves high long-term reward [56, 41]. One of the key challenges of the RL setting is that credit assignment, exploration, and generalization [56] must be addressed even when the agent

has seen very little data and thus has low quality representations [22, 11]. When the representations are low quality, determining a desirable state to reach and finding a policy to reach that state are both difficult [17]. Intuitively, learning a compressed representation should help to address these challenges. If extraneous information can be removed, it should be easier to generalize to new samples from the environment.

Approaches from the RL theory literature have shown benefits from compressed representations in the discrete latent state setting [40], [10], [8], [63]. The HOMER algorithm [40] explores by trying to reach the frontier of pairs of the discrete latent states and actions with the lowest counts. While these algorithms give strong theoretical guarantees [9], planning and exploring with them does not scale beyond a small number of discrete states.

We explore the intersection between theoretically-grounded representation learning in small tabular-MDPs and representations for the deep reinforcement learning setting. We seek to retain the expressiveness of factorial representations while making the representation compressed [35, 34]. In our proposed method (Figure 1), Representations for RL with Discrete Information Bottleneck (REPDIB), we make the representations discrete and factorial, while also encouraging them to be parsimonious through a gaussian variational information bottleneck [2, 13, 14, 12]. These are expressive enough to model complicated environments, yet avoid the unbounded complexity of unstructured continuous representations.

This work studies the effectiveness of learning compressed representations for reinforcement learning. We find that by using an information bottleneck that induces a factorial structure in the embedding space, REPDIB can learn more robust representations. This improvement is especially pronounced in settings where the observation con-

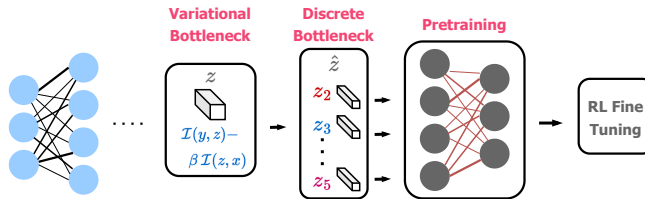


Figure 1: **Illustration** of the generic approach of REPDIB, where we learn representations with variational and discrete factorial bottlenecks. We show that pre-training Representations with Discrete Information Bottleneck (REPDIB) leads to learning of robust representations, especially when observations consist of irrelevant and exogenous information

tains exogenous noise [10, 9], which is any information that is unrelated to the agent’s actions. We propose an easy to use approach effective for improving downstream performance in settings with irrelevant background information. Our work offers the following contributions: (a) Learning representations which more closely match the salient attributes of the environment, and improved robustness by learning factorial representations that can ignore irrelevant information in a practical robot arm task (b) Improved sample efficiency due to structured representations, for better generalization in continuous control (c) bottleneck representations that can improve robustness in offline RL in presence of exogenous distractors. Through range of experiments, we show that REPDIB learns compressed representations, which helps in exploration and reward-free pre-training of representations to improve efficiency and robustness on downstream tasks.

2 Related Work

Self Supervised Representation Learning in RL. Several prior works have studied representation learning in context of RL, ranging from online to offline settings [64, 26, 43], while also studying the ability to recover underlying latent states to capture environment dynamics [28, 4]. Most of these works involve learning representations from high dimensional observations, which may contain irrelevant information. This is formalized as learning under irrelevant exogenous information [10, 9], by the theoretical RL community studying representation learning. In this work, we show effectiveness of information bottlenecks with REPDIB, when learning under exogenous information, and show that bottlenecks can filter out irrelevant information from observations. Empirically, prior works studied regularized objectives, for learning robust representations [39, 21] while others have exploited empowerment based objectives [42]. Self supervised objectives, when used for pre-training representations have shown to achieve tremendous performance improvements [29, 54, 51, 52], while when learning with fine-tuning representations, it leads to better exploratory objectives [65].

Learning Minimal Representations with Information Bottleneck. In this work, we argue that information bottleneck based representations with REPDIB can be an ef-

fective approach for learning robust representations in RL, in presence of exogenous information. The information bottleneck principle [61, 58, 53, 57] advocates for learning minimal sufficient representations, i.e. those which contain *only* sufficient information for the downstream task. Optimal representations contain relevant information between X and Y that is parsimonious to learn a task. Several approaches have been proposed to design information bottlenecks in deep learning models, such as variational bottlenecks [55, 2] and discrete representation bottlenecks [59]. Most prominently, Alemi et al. [2] introduced a variational approximation to a mutual information objective of the information bottleneck and applied this to deep neural networks.

Information Bottleneck for Exploration in Deep Reinforcement Learning. The exploration problem is inherently coupled with the representation learning problem, since discovering underlying latent structure of the world ensures that the agent learns about the unseen frontiers in observation space to reach. While several recent works have studied representation learning in RL for improving downstream task performance [51, 52], the closest to our work is learning with prototypical representations [65], which studies the coupled problem of representation learning and exploration. [13, 12] previously studied exploration based on identifying latent bottleneck states, but do not learn an explicit representation with a self-supervised objective. [24] studied bottlenecks for inducing exploration in RL. On the theoretical side, [40] grounds representation learning and exploration with theoretical guarantees, but cannot scale to rich observation environments. Several prior works in exploration have been proposed, with large observation spaces, such as using pseudo-counts [46, 5], optimism-driven exploration [45], intrinsic motivation [47], random network distillation [6] and curiosity based exploration with prediction errors [48]. While these algorithms propose exploration in complex high dimensional tasks, they do not necessarily learn and exploit any form of structure in the representation space.

Comparisons with Prior Related Works : An information bottleneck aiming at minimal sufficient representations can be implemented in various ways, including a variational approach (VIB) and architectural choices such

as reducing the dimension of deeper layers, or by discretizing layers. Chenjia et al. [3]. directly apply the information bottleneck to the dynamics of the system, whereas RepDIB applies it for different downstream targets, such as DQN targets or inverse model targets. RepDIB also combines both kinds of bottlenecks, i.e. architectural (discrete bottlenecks in particular) and variational ones. Previous work in reinforcement learning which enforce bottlenecks have worked with either type independently. Dreamer-v2 and similar variants have included discretization for pixel-level model-based learning. In this paper, we take a zoomed-out perspective on the efficacy of bottlenecks in learning representations for reinforcement learning.

3 Discrete Factorial Information Bottlenecks in Representation Learning

The goal of this work is to study the effectiveness of variational and discrete information bottlenecks in representation learning. While several prior works have studied representation learning for RL, we show that especially when observations can contain irrelevant information, addition of simple bottlenecks can lead to learning effective robust representations for improving performance on downstream tasks. Through a range of experiments, as in section 4, we show that RepDIB learns a structured representation space, via use of discrete information bottlenecks [34], that can be quite effective for downstream learning. In this section, we briefly describe our approach for learning robust representations with information bottlenecks.

The RepDIB technique begins with a hidden representation $\mathbf{z} \in \mathbb{R}^m$ for a rich-observation x . This \mathbf{z} could be the output of a convolutional neural network, a recurrent neural network, transformer, or any other expressive neural model. We induce a compositional structure in the learnt representation space by using a vector quantization discretization bottleneck [60]. This is achieved by using a discretization module with G factors each with L codes. Thus the total number of discrete states that we can express is L^G . We can learn embeddings using multiple G factors and can concatenate them into a single embedding $\hat{\mathbf{z}} = \phi(\mathbf{z})$ with $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^m$. Thus the discretization bottleneck ϕ preserves the size of the hidden representation.

While the compositional structure can solely be achieved through the discretization bottleneck, we additionally add a gaussian information bottleneck [2]. This is added directly before the discretization function ϕ . encourage more parsimonious discrete representations. Adding an information bottleneck to capture sufficient representations means that the we can achieve better compositionality by using *fewer* discrete codes. Figure 3 shows the learnt compositional structure in the latent embedding space extracted by RepDIB, while no apparent structure exists in the latent space for a baseline without any bottleneck. Following the learnt embeddings, we then apply the

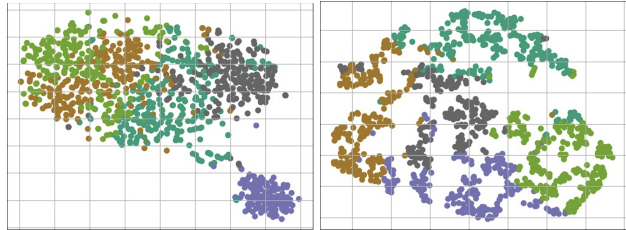


Figure 2: **T-SNE analysis** comparing representation embeddings. We take the ProtoRL [65] setup for learning representations in continuous control RL, where RepDIBbased information bottlenecks are applied on top of the learnt representations from ProtoRL. **Left.** Latent representations from ProtoRL with discrete prototypes. **Right.** Factorized latent representations with ProtoRL + RepDIB, that learns better structure in the representation space, when we apply a variational (Gaussian) information bottleneck followed by discrete information bottlenecks.

VQ discretization bottleneck, with different grouping factors. To apply discretization bottleneck, we quantize the output of the projector layer into group-based discrete latent embedding. Concretely, instead of assigning each continuous embedding \mathbf{z}_e to a single one discrete vector, we first divide each continuous state representation into G different groups as $\mathbf{z}_e = \text{concat}(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_G)$, then we assign each segment $\mathbf{c}_i \in \mathbb{R}^{\frac{m}{G}}$ separately to a discrete vector $\mathbf{e} \in \mathbb{R}^{L \times \frac{m}{G}}$ using a nearest neighbour look-up: $\mathbf{e}_{\mathbf{o}_i} = \text{DISCRETIZE}(\mathbf{c}_i)$, where $\mathbf{o}_i = \text{argmin}_{j \in \{1, \dots, L\}} \|\mathbf{c}_i - \mathbf{e}_j\|$, where L is the size of the discrete latent space (i.e., an L -way categorical variable). After that, we concatenate all segments to obtain the discrete embedding $\mathbf{z}_q = \text{CONCATENATE}(\text{DISCRETIZE}(\mathbf{c}_1), \dots, \text{DISCRETIZE}(\mathbf{c}_G))$. This process results in compositionality of latent representation with an information bottleneck.

RepDIB Implementation Details. We provide technical details of how our approach can be implemented on any existing self-supervised reinforcement learning objective (Figure 1). To enable factorial structure in the representation space, we can integrate a vector quantization discretization bottleneck on top of any encoder that learns a latent state representation. Given an encoder that maps observations o to latent representation $\phi(\cdot)$, we first use a variational information bottleneck (VIB) based on reparameterization, with a uniform Gaussian prior. We then quantize the continuous representation from an information bottleneck into discrete latent variables, generalizing vector quantization in VQ-VAE.

4 Experiments : Representations with Information Bottleneck

We seek to understand the effectiveness of information compression in representations. We emphasize that RepDIB can be applied as a plug-in approach, on top of any existing framework that learns representations with a

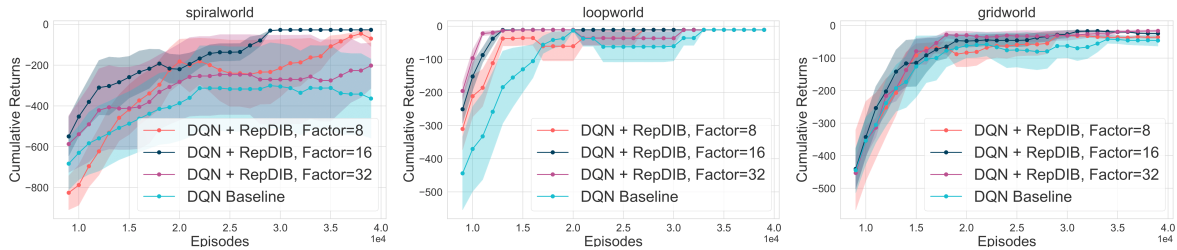


Figure 3: **Performance** comparison on 3 different maze navigation tasks, with REPDIB, using different factors 8, 16, 32 in the learnt representation, integrated on a baseline DQN agent.

self-supervised objective. Through our experiments, we answer the following questions:

Does inducing structure in representation space help with exploration? We first demonstrate on simple toy tasks that learning representations with REPDIB can induce a factorized structure that can lead to effective exploration. By using discrete information bottlenecks, we can recover the underlying discrete latent states while also learning factorized embedding space, that leads to better exploration in maze tasks when using a simple DQN agent.

Does factorized representations with REPDIB help learn task agnostic pre-trained representations, for better generalization capabilities? We evaluate REPDIB on several complex control tasks using the URLB benchmark [30] for testing generalization capabilities. In this setting, we pre-train representations in a reward-free approach on a given task, followed by fine-tuning on different downstream tasks. Most importantly, we show that sample efficiency of REPDIB can further be improved as a function of pre-training steps, where REPDIB can improve downstream performance with only few pre-training steps.

Does information bottleneck help learn parsimonious representations to learn relevant representations in a real robot arm task, while ignoring background distractors? To answer this, we use a real robot arm collected data, with a temporal background structure, where there is background noise from lightnings, TV and video. In this setting, we show that REPDIB can capture the relevant factors of variation and ignore irrelevant distractors through the use of information and discretization bottleneck.

Does bottleneck representations help in sequence modelling problem from offline datasets? We evaluate REPDIB in the offline Atari benchmark where environment observations consist of additional exogenous information, using the Decision Transformer [7], and find that pre-trained representations with REPDIB can learn robust representations ignoring distractors.

What is the impact of VQ information bottleneck to extract unimodal and fuse multi-modal representations? We study REPDIB using existing human activity recogni-

tion based dataset in a multi-modal learning setting. We show that when fusing representations from single modality with information bottleneck, followed by compressing the resulting multi-modal representation, REPDIB helps achieve performance improvements compared to existing baselines with information bottleneck on multi-modal representations for activity recognition tasks.

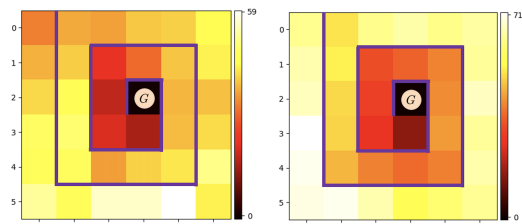


Figure 4: **Relative Distance in Representation Space.** (Left): Baseline DQN Agent (Right): DQN Agent with REPDIB. We show that REPDIB learns representations that can better capture underlying topology of how the agent can move in the maze. Darkness of position shows how similar the representation is to the point in the center.

4.1 Maze Navigation Tasks

Experiment Details. We use three kinds of maze navigation tasks to evaluate the effectiveness of learning parsimonious representations with REPDIB: *GridWorld*, *SpiralWorld*, *LoopWorld*. We first learn the state representations on *GridWorld* with data collected by a random policy, and then adapt the pre-trained representations to all these three tasks to learn the end-task policy.

Experiment Results. We study spiral and loop world maze navigation task, with a baseline DQN agent. During fine-tuning based on pre-trained representations from an empty gridworld, we simultaneously update both the representations and the DQN agent, given pixel based observations. We find that the induced factorized representation structure leads to better coverage of the state space, as demonstrated in Figure 6 while also capturing topology of the maze in representation space as shown in figure 4. For the self-supervised objective for learning representations, we use

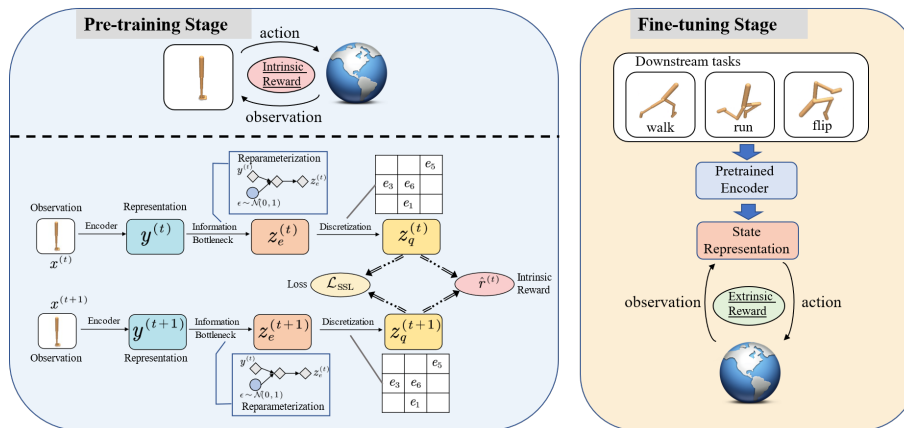


Figure 5: **Summary of REPDIB** integrated on top of the ProtoRL baseline [65] for testing generalization capability in continuous control tasks from URLB benchmark [30]. We find that REPDIB improves intrinsically-motivated exploration (left). An information bottleneck is used to encourage the discrete codes to be parsimonious. The reward-based fine-tuning stage remains unchanged when using REPDIB (right).

DRIML [39] for reward-free representation learning, followed by REPDIB integrated on top of the encoder. Experiment results, as in figure 3 shows that the induced structure, based on different group factors of the discretization bottleneck, leads to improved performance compared to a baseline DQN agent.

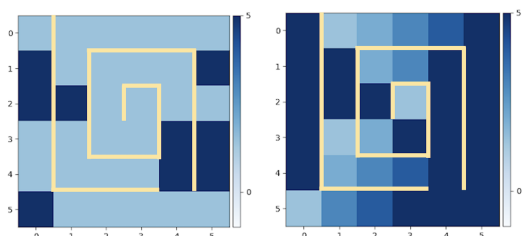


Figure 6: **State Space Coverage** comparing DQN agent (left) and DQN agent with REPDIB (right). Factorized Representations with REPDIB leads to better state space coverage in reward-free exploration.

4.2 Generalization in Continuous Control

REPDIB is then evaluated on a range of continuous control tasks with visual observations. We integrate REPDIB on the state of the art Proto-RL baseline [65], as shown in figure 5 which has been shown to learn good representations from pre-training, for better fine-tuning performance. We follow the experiment setup from the URLB benchmark [30], and explain our experiment setup below, comparing REPDIB with a baseline Proto-RL agent, since it has been shown to outperform other baselines learning self-supervised representations.

Experiment Details. The key to Proto-RL is to learn a set of prototypical vectors and prototypes by projecting the embeddings onto clusters, referred to as prototypes. To ensure exploration, a latent state entropy distribution is opti-

mized with an approximation, based on the learnt prototypical representations. This can form the basis for an intrinsic reward function, which ensures sufficient coverage in a pre-training phase in a task agnostic reward-free setting. In contrast to Proto-RL, the key to our approach of learning discrete prototypes, REPDIB is to ensure that a factorized structure is learnt in the latent representation. We refer to this as *structured exploration*, since REPDIB exploits the use of information bottleneck and vector quantization to induce a factorial structure embeddings.

We use a total of 12 continuous control tasks with varying difficulty (3 different domains with 4 different downstream tasks per domain): *Walker*, *Quadruped* and *Jaco Arm*. The agent is pre-trained on a specific task in a given domain, and then adapts to the other downstream tasks within that domain. We follow the same experiment pipeline as in URLB benchmark [30]. We checkpoint the agent at 100k, 500k, 1M, 2M time-steps during pre-training, and then evaluate the adaptation ability of the method by adapting the pre-trained policy to downstream tasks.

Pre-Training: During the pre-training phase, we train REPDIB agent in a task agnostic reward free setting. The goal here is to encourage agent to reach unseen regions to collect more diverse data, such that this can further help in learning better representations. For this, we follow a similar procedure as Proto-RL, where the agent is trained to maximize coverage by estimating an approximation to the entropy of the latent state distribution [65]. In case of REPDIB, instead of estimating entropy based on an unstructured representation, REPDIB utilize the factorization structure in the representation space, through the use of the information bottleneck followed by the discretization module. Therefore, REPDIB computes the intrinsic reward based on the discrete factorial embeddings for more efficient structured exploration.

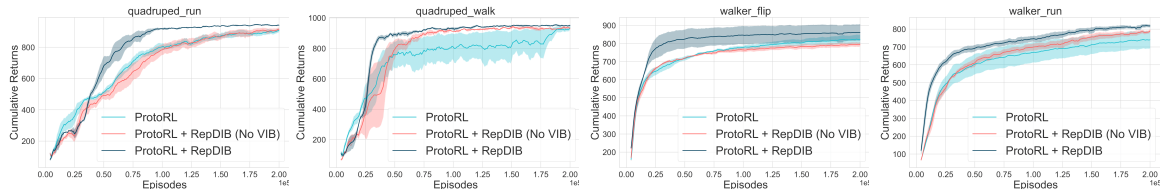


Figure 7: **Fine-Tuning performance** on different domains, with pre-trained representations learnt with REPDIB and comparison with ProtoRL baseline.

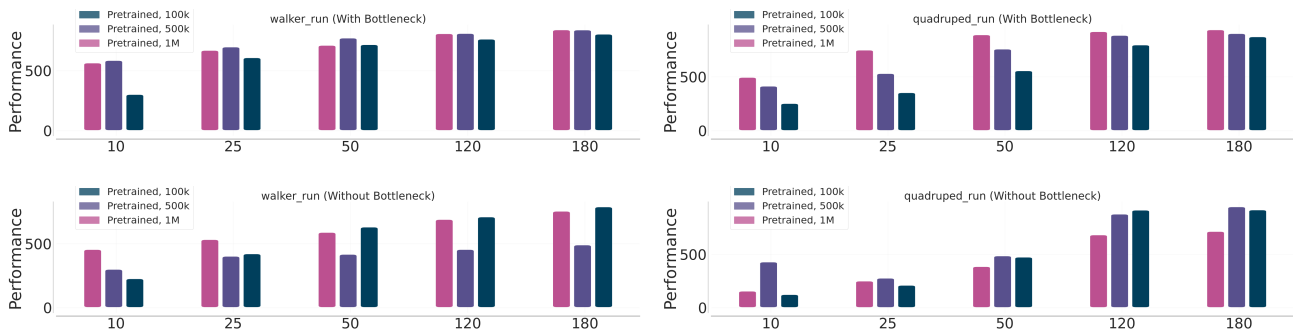


Figure 8: **Downstream impact** of varying the number of pre-training steps (100K, 500K or 1M timesteps). X-axis shows different fine-tuning steps. We see that with the use of VIB and a discretization bottleneck, there is a gradual improvement in performance (**top row**); however the performance during fine-tuning can degrade as a function of pre-training steps when a REPDIB based bottleneck is not used in the baseline Proto-RL agent (**bottom row**).

Fine-Tuning. In the second phase, the agent is fine-tuned to solve new tasks to test its generalization capability. We use the learnt representation to then collect a dataset, which is then used by any standard state based off-policy RL algorithm such as soft actor critic (SAC) [15]. Since the compositional structure is mostly exploited during pre-training phase, during fine-tuning we freeze the learnt representation to study the effectiveness of reward free representation.

Experiment Results on Control Tasks. Since Proto-RL is shown to outperform existing baselines, including random exploration DrQ [66], curiosity based exploration ICM [49] and unsupervised active pre-training (APT) [36]; in this work we mostly compare to the state of the art Proto-RL baseline. We provide comparisons for REPDIB including the variational information bottleneck and REPDIB which only includes the discretization bottleneck (denoted as REPDIB only). Having pre-trained the representation encoder with and without a bottleneck in reward-free setting, we test the fine-tuning performance of the RL algorithm based on the fixed and learnt representation. Figure 7 demonstrates the significance of REPDIB algorithm. We find that the use of information bottleneck prior to the discretization, can significantly help to improve sample efficiency during fine-tuning. We further examine the significance of the information bottleneck with different KL weightings in Appendix figure 28

Fine-Tuning Performance as a Function of Pre-Trained Unsupervised Representation Steps. Downstream per-

formance should monotonically improve with more steps of pre-training. However, it has been found that downstream performance sometimes degrades with more pre-training steps and that this counter-intuitive failure mode is common to all of the most widely used unsupervised RL algorithms [30]. We reproduced this phenomenon in our prototypical-RL baseline. We found that REPDIB alleviates this problem, resulting in monotonic improvements in downstream performance with more pre-training steps (Figure 8).

4.3 Offline Experiments with Exogenous Distractors

Experiment Setup with Atari using Decision Transformer. We first consider the reward-conditioned behavior cloning setup with decision transformers [7], where the goal is to learn representations that can ignore noisy or background information not relevant to the task using REPDIB. We consider the 4 games considered in [7] (Pong, Breakout, Seaquest, Qbert), using offline dataset [1] for training. The model is trained using a sequence modeling objective to predict the next action given the past states, actions, and returns-to-go.

To add exogenous information to the observation space, we append a randomly sampled cifar10 [27] image to each frame. We keep the cifar image fixed in an episode but use a different image across episodes. We first pretrain our convolutional encoder with multi-step inverse objective introduced in [28]. We then train the Decision Transformer for

GAME	MULTI-STEP INVERSE	MULTI-STEP INVERSE + REPDIB
PONG	11.4 \pm 2.653	12.8 \pm 2.561
QBERT	878.6 \pm 745.146	1100.0 \pm 898.499
BREAKOUT	19.8 \pm 3.059	41.8 \pm 7.305
SEQUEST	915.2 \pm 126.368	1058.4 \pm 116.629

Table 1: **Atari Results.** We compare the proposed MULTI-STEP INVERSE + REPDIB to MULTI-STEP INVERSE on 4 atari game using the Decision Transformer setup. We can see that the proposed approach outperforms the baseline in all cases. Results averaged across 5 seeds.

action prediction keeping the convolutional encoder fixed. For the proposed approach, we discretize the output the encoder as described in method section 3 before applying multi-step inverse objective.

Experiment Results. Table 1 summarizes the Atari results, where MULTI-STEP INVERSE + REPDIB outperforms MULTI-STEP INVERSE in all games thus showing the effectiveness of the VQ bottleneck. We use a discretization module with 32 factors for all the games. Additional results analysing the effect of the number of discretization factors is presented in appendix.

Experiments with Visual Offline RL. We then consider the visual pixel-based offline dataset [38] for control, where we learn representations using a MULTI-STEP INVERSE model [28]. We consider two settings : with no visual background distractors and another where we add time correlated exogenous image distractors in the background. Figure 9 summarizes the results, where we find that in presence of exogenous image distractors, REPDIB can learn more robust representations during pre-training; whereas performance is similar in the setting without any additional exogenous distractors.

Comparisons with Other Information Bottleneck Approaches : We now compare REPDIB with several other bottleneck baselines in the pixel based offline RL setup. We follow the same experiment setup as described in section 4.3 and integrate information bottleneck approaches on top of three existing representation learning objectives, namely AC-State [28], One step inverse dynamics [48] and DRIML [39]. We compare with **three different baselines** along with comparisons of variations of REPDIB bottleneck.

Note that the other baselines we compare with are all based on approximations of a mutual information based objective. In contrast, REPDIB does not require any MI based approximations. We mainly compare with EMI(with MINE objectives)[23], DB(Dynamic Bottleneck) [3] and SVIB [67, 50] as reviewers have pointed out, and show in figures 10 and 11 how REPDIB compares with other baselines. Specifically, EMI proposes to maximize the mutual information of state embedding representations and action embedding representations by maximizing the estimated

lower bounds of both mutual information. DB follows the Information Bottleneck principle to learn dynamics-relevant representation by maximizing the mutual information $I(Z_t; S_{t+1})$ while minimizing the mutual information $I([S_t, A_t]; Z_t)$, where Z_t is a compressed latent representation of (S_t, A_t) , S_t, A_t correspond to the current state and current action respectively. SVIB utilizes the mutual information between the observation and its corresponding representation as an additionally penalized term for standard loss function in RL, optimizing all networks by Stein Variational Gradient Descent (SVGD). Notably, for a fair comparison with the other bottlenecks, we update all networks by just using Adam optimizer instead of SVGD. We emphasize that compared to the baselines, REPDIB is easy to integrate since it only requires adding a VQ-VAE based factorization with a variational information bottleneck.

4.4 Robot Arm Experiment in Presence of Irrelevant Background Information

Experiment Details. We then evaluate REPDIB on a challenging real robot dataset, containing high resolution video of a robot arm in presence of a rich temporal background noise [28]. For learning a latent state representation of the images, we use a multi-step inverse model [28, 10], and integrate REPDIB with the information and discretization bottleneck on the learnt representation. In this task, the robot arm moves on top of a grid layout, containing 9 different positions. We denote these as the *true states*. We collect a dataset containing pixel based observations only, where the images consist of the robot arm along with the background distractors. Inspired by the exogenous noise information setup [10], we setup the robot task while there is a TV playing a video in the background, with other flashing lights nearby. The offline dataset consists of 6 hours of robot data, with 14000 samples from the arm, taking high level actions of move left, right, up and down. A sample point image is collected after each action, and the background distractors changes significantly, due to video and lighting in the background. The goal of the experiment is to predict accurately the ground truth state position by learning latent representations with REPDIB.

Experiment Results. We evaluate the ability of REPDIB to accurately reconstruct the image, by learning the latent state representation while also ignoring the background distractors. This is denoted as the *Image Noise*, where we compare REPDIB with and without VIB, alongside a baseline agent which only learns a representation. For learning latent representations, we use a multi-step inverse dynamics model [10]. In addition, we compare the ability of REPDIB to accurately predict the ground truth states, denoted by *State Accuracy* solely from the observations, as a classification task. This is challenging since the learnt representation needs to predict ground states while ignoring the irrelevant background information. Furthermore with

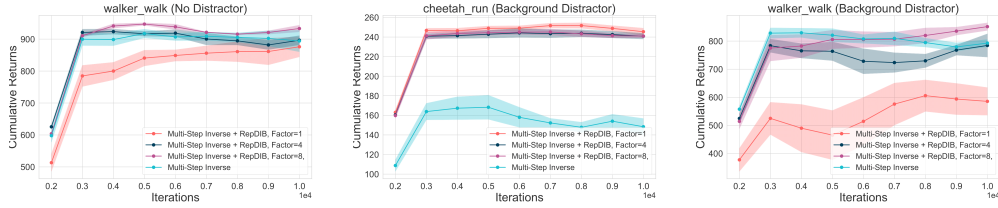


Figure 9: **REPDIB can learn more robust representations** due to information bottleneck, in presence of background exogenous distractors, when using the offline visual control setup from [38]. In contrast, performance is almost similar in settings without distractors.

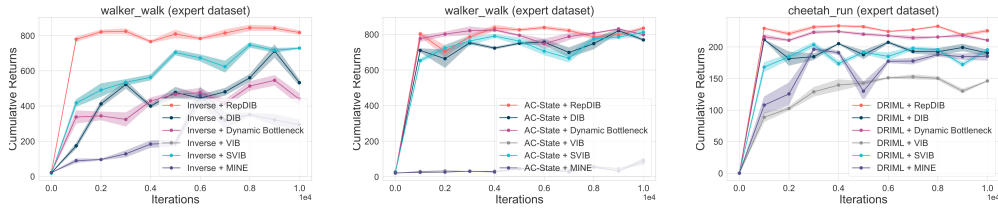


Figure 10: **Time correlated exogenous images in the background.** Comparison of REPDIB with other approaches based on information bottleneck based approximations in the offline RL setup. Following our previous results in the main, we now compare different bottleneck based approaches on top of existing representation learning objectives.

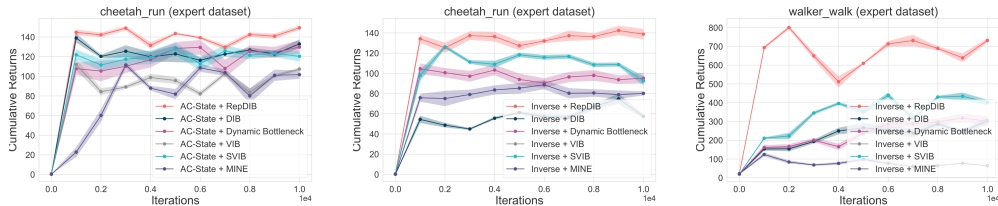


Figure 11: **Changing video distractors, as exogenous noise in the background.** We now consider a slightly more difficult setup where we have changing video distractors in the background. We again compare REPDIB with other information bottleneck based approaches, when integrated on top of existing representation learning objectives


the learnt model, we predict the time-step for each observation as an additional metric to determine effectiveness of REPDIB. The time-step is an indicator of the background noise that appeared in each sample; and with *Temporal Noise*, we evaluate REPDIB to predict the time step while ignoring irrelevant information from observations. Experiment results in Figure 12 shows that the use of VIB helps improve the ability of REPDIB to remove noise from the representation, while being able to almost perfectly predict the ground truth state of the robot.

4.5 Multi-Modal Representation Learning with Information Bottleneck

Experiment Details. We evaluated the impact of REPDIB to learning multi-modal representations for human activity recognition task. We extended the baseline multi-modal models in two ways to incorporate VQ bottleneck: **REPDIB+MM:** We extract multi-modal representations

using existing models (e.g. Keyless [37] and HAMLET [18]) and then apply VQ bottleneck on the fused multi-modal representations. **REPDIB+MM(REPDIB+Uni):** We applied VQ bottleneck in two steps. First, we extract unimodal representations and apply VQ bottleneck to produce discretized unimodal representations. These discretized representations are fused and passed through a VQ bottleneck to produce task representations for the activity recognition. In the baselines, we used five modalities: two viewpoints of RGB videos and three wearable sensors (acceleration, gyroscope, and orientation). We evaluated all the baselines on the MMAcT dataset in a cross-subject evaluation setting and reported F1-Score of activity recognition task [25].

Experiment Results. The results in Table 2 suggest that applying VQ bottleneck on the multi-modal representations degrades the performance of multi-modal models for the activity recognition task. For example, applying VQ bottle-



Bottleneck	None	REPDIB(No VIB)	REPDIB
Temporal Noise Relative Error	1.0	0.7043	0.6650
Visual Noise Relative Error	1.0	0.9725	0.9713
State Estimation Relative Error	1.0	1.001	0.9988

Figure 12: **Representations learned from videos of a real robotic arm** (with various distractors such as a TV and color-changing lights). We evaluate the representation quality with various types of bottlenecks. REPDIB is best able to remove noise from the representation without removing information about the true state of the robot

Method	F1-Score (%)
SMD [16]	63.89
Multi-Teachers [25]	62.67
MMAD [25]	66.45
HAMLET [18]	69.35
Keyless [37]	71.83
REPDIB+MM(HAMLET)	57.47
REPDIB+MM(Keyless)	63.22
REPDIB+MM(REPDIB+Uni)	69.39

Table 2: **Cross-subject performance** comparison (F1-Score) of multi-modal learning model on MMAAct dataset

neck on multi-modal representations from Keyless model (REPDIB + $MM(Keyless)$) significantly degrades the F1-Score of the activity recognition task to 57.47% from 69.35%. In these models, non-discretized unimodal representations are fused to produce a compressed and non-discretized multi-modal representation. The results suggest that applying VQ bottleneck on non-discretized multi-modal representation can not ensure retaining salient representations for task learning.

On the other hand, applying VQ bottleneck both on the unimodal and multi-modal representations improves the performance of the models compared to the models that do not use REPDIB or use REPDIB only on the multi-modal representations. For example, REPDIB + $MM(REPDIB + Uni)$ model uses the same HAMLET model and applies REPDIB on the unimodal and multi-modal representations. REPDIB + $MM(REPDIB + Uni)$ slightly improves the performance of HAMLET. REPDIB + $MM(REPDIB + Uni)$ fuses the discretized unimodal representations using a modality weighting approach, which is modeled as 1D-CNN. As several works on multi-modal representation learning showed that the way to fuse unimodal representations could impact the performance of the downstream task [19, 20, 32], there is room for improvement by effectively fusing the discretized unimodal representations. Moreover, as a couple of hyper-parameters in VQ bottle-

neck impact the model performance, such as the number of groups and number of embeddings, finding the appropriate value of hyper-parameters can improve the model performance. Thus, our experimental results show a crucial future avenue of research to utilize REPDIB information bottleneck for extracting salient multi-modal representations.

5 Discussion

Conclusion. Representation learning methods in RL have been extensively studied in the recent past. However, when learning directly from observations consisting of exogenous information, the need for learning robust representations becomes vital. To this end, we propose REPDIB that learns robust representations by inducing a factorized structure in embedding space. Our work shows that discrete bottleneck representations that compress the relevant information from observations, can lead to substantial improvements in downstream tasks, as shown in our experimental results.

Limitations and Future Work. Whether the bottlenecks with different factors *truly* lead to a compositional representation space that can disentangle different factors of observations is an interesting avenue for future work. While we enforce discrete factorization, we provide no theoretical proof that this corresponds to an actual factorization structure in the data. We believe that inducing such compositional structure can shape the path towards truly achieving better generalization capabilities of RL agents. How to achieve and leverage a compositional representation space for better generalization remains an interesting question, both theoretically and empirically.

Acknowledgements

The authors would like to thank Remi Tachet Des Combes, Romain Laroche, Harm Van Seijen, and Doina Precup for valuable feedback on the draft. Hongyu Zang and Xin Li were partially supported by NSFC under Grant 62276024 and 92270125.

References

- [1] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.
- [2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.
- [3] Chenjia Bai, Lingxiao Wang, Lei Han, Animesh Garg, Jianye Hao, Peng Liu, and Zhaoran Wang. Dynamic bottleneck for robust self-supervised exploration. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17007–17020, 2021.
- [4] Philip J. Ball, Cong Lu, Jack Parker-Holder, and Stephen J. Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 619–629. PMLR, 2021.
- [5] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1471–1479, 2016.
- [6] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.
- [7] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097. Curran Associates, Inc., 2021.
- [8] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 09–15 Jun 2019.
- [9] Yonathan Efroni, Dylan Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*. PMLR, 2022.
- [10] Yonathan Efroni, Dipendra Kumar Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provable rl with exogenous distractors via multistep inverse dynamics. *ArXiv*, abs/2110.08847, 2021.
- [11] Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *CoRR*, abs/2112.13487, 2021.
- [12] Anirudh Goyal, Yoshua Bengio, Matthew Botvinick, and Sergey Levine. The variational bandwidth bottleneck: Stochastic evaluation on an information budget. *arXiv preprint arXiv:2004.11935*, 2020.
- [13] Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew M. Botvinick, Yoshua Bengio, and Sergey Levine. Infobot: Transfer and exploration via the information bottleneck. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [14] Anirudh Goyal, Shagun Sodhani, Jonathan Binas, Xue Bin Peng, Sergey Levine, and Yoshua Bengio. Reinforcement learning with competitive ensembles of information-constrained primitives. *arXiv preprint arXiv:1906.10667*, 2019.
- [15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865. PMLR, 2018.
- [16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS*, 2015.
- [17] Baihe Huang, Kaixuan Huang, Sham M. Kakade, Jason D. Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Going beyond linear RL: sample efficient neural function approximation. In Marc’Aurelio Ranzato,

- Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8968–8983, 2021.
- [18] Md Mofijul Islam and Tariq Iqbal. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10285–10292, 2020.
- [19] Md Mofijul Islam and Tariq Iqbal. Mumu: Cooperative multitask learning-based guided multimodal fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):1043–1051, Jun. 2022.
- [20] Md Mofijul Islam, Mohammad Samin Yasar, and Tariq Iqbal. MAVEN: A memory augmented recurrent approach for multimodal fusion. In *IEEE Transaction on Multimedia*, 2022.
- [21] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [22] Sham M. Kakade. On the sample complexity of reinforcement learning. Phd thesis, University College London, 2003.
- [23] Hyoungseok Kim, Jaekyeom Kim, Yeonwoo Jeong, Sergey Levine, and Hyun Oh Song. Emi: Exploration with mutual information. In *International Conference on Machine Learning (ICML)*, 2019.
- [24] Jaekyeom Kim, Minjung Kim, Dongyeon Woo, and Gunhee Kim. Drop-bottleneck: Learning discrete compressed representation for noise-robust exploration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [25] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klunkigt, Bin Tong, and Tomokazu Murakami. MMAct: A large-scale dataset for cross modal human action understanding. In *ICCV*, pages 8658–8667, 2019.
- [26] Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5774–5783. PMLR, 2021.
- [27] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.
- [28] Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Lekan Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of controllable latent states with multi-step inverse models. *arXiv preprint arXiv:2207.08229*, 2022.
- [29] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 2020.
- [30] Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: unsupervised reinforcement learning benchmark. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.
- [31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5543–5551. IEEE Computer Society, 2017.
- [32] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.
- [33] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [34] Dianbo Liu, Alex Lamb, Xu Ji, Pascal Not-sawo, Mike Mozer, Yoshua Bengio, and Kenji Kawaguchi. Adaptive discrete communication bottlenecks with dynamic vector quantization. *arXiv preprint arXiv:2202.01334*, 2022.

- [35] Dianbo Liu, Alex M Lamb, Kenji Kawaguchi, Anirudh Goyal ALIAS PARTH GOYAL, Chen Sun, Michael C Mozer, and Yoshua Bengio. Discrete-valued neural communication. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2109–2121. Curran Associates, Inc., 2021.
- [36] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18459–18473, 2021.
- [37] Xiang Long, Chuang Gan, Gerard De Melo, Xiao Liu, Yandong Li, Fu Li, and Shilei Wen. Multimodal key-less attention fusion for video classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [38] Cong Lu, Philip J. Ball, Tim G. J. Rudner, Jack Parker-Holder, Michael A. Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations. *CoRR*, abs/2206.04779, 2022.
- [39] Bogdan Mazouze, Remi Tachet des Combes, Thang Doan, Philip Bachman, and R. Devon Hjelm. Deep reinforcement and infomax learning. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- [40] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.
- [41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [42] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.
- [43] Ofir Nachum and Mengjiao Yang. Provable representation learning for imitation with contrastive fourier features. *CoRR*, abs/2105.12272, 2021.
- [44] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *WACV*, pages 53–60. IEEE, 2013.
- [45] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20:124:1–124:62, 2019.
- [46] Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR, 2017.
- [47] Pierre-Yves Oudeyer and Frédéric Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers Neurobotics*, 1:6, 2007.
- [48] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.
- [49] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2778–2787. PMLR, 2017.
- [50] Yingjun Pei and Xinwen Hou. Learning representations in reinforcement learning: An information bottleneck approach. *CoRR*, abs/1911.05695, 2019.
- [51] Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron C. Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [52] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. In Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy

- Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12686–12699, 2021.
- [53] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.
- [54] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 2021.
- [55] Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and S Yu Philip. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4165–4174, 2022.
- [56] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [57] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- [58] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (itw)*, pages 1–5. IEEE, 2015.
- [59] Frederik Träuble, Anirudh Goyal, Nasim Rahaman, Michael Mozer, Kenji Kawaguchi, Yoshua Bengio, and Bernhard Schölkopf. Discrete key-value bottleneck, 2022.
- [60] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [61] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.
- [62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
- [63] Zhihan Xiong, Ruoqi Shen, and Simon S Du. Randomized exploration is near-optimal for tabular mdp. *arXiv preprint arXiv:2102.09703*, 2021.
- [64] Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11784–11794. PMLR, 2021.
- [65] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021.
- [66] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [67] Pei Yingjun, Hou Xinwen, Li Jian, and Lei Wang. Optimizing information bottleneck in reinforcement learning: A stein variational approach. 2021.

Supplementary Materials

Appendix

6 Additional Experiment Results and Details

6.1 Visual Offline RL with Exogenous Observations in Datasets

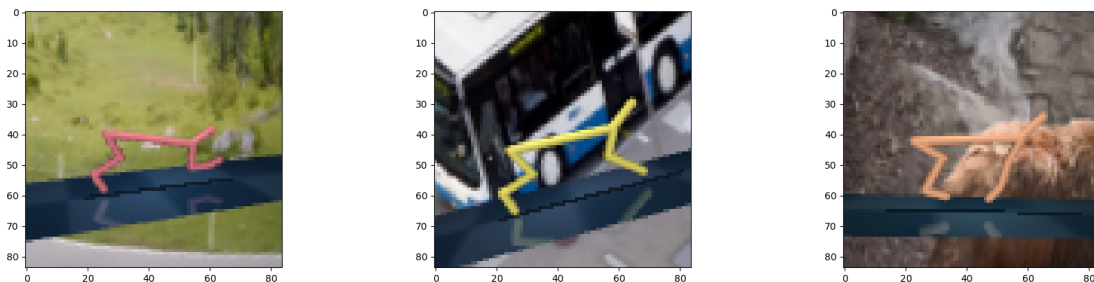


Figure 13: Sample observations from the visual offline datasets with exogenous time correlated images in the background. The exogenous background image changes per episode during offline data collection

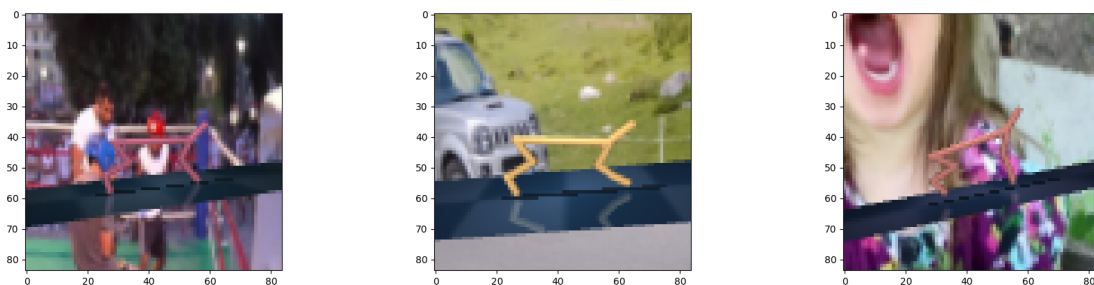


Figure 14: Sample observations from the visual offline datasets with exogenous changing video in the background. The exogenous background video distractor changes per episode during offline data collection

Experiment Setup and Details : We evaluate REPDIB on visual offline datasets from the v-d4rl benchmark [38]. Data collection details on different domains are provided in [38]. In addition, we also consider an extension of the v-d4rl benchmark, where we re-collect the data with additional *exogenous noise* present in the observations. We follow the same data collection procedure as in v-d4rl, except during data collection, there are two variations of exogenous noise that is considered. 1. We first consider a time correlated exogenous noise setting where during data collection, at each episode the agent sees the environment observation and an additional background image from the CIFAR dataset. This image changes per episode of data collection, and we introduce this such that learning robust representations by avoiding the distractors plays an important role for policy learning. 2. We then consider a setting where instead of images that changes per episode, we now have a video distractor that changes at every episode of data collection. This is considered an even harder setting since the agent sees the observations while in addition there are unrelated video data playing in background.

For the baseline policy optimization RL algorithms, we follow the same experiment pipeline as in [38]. The major difference being, we additionally train the encoders with a representation learning objective where we pre-train the encoders with a fixed $100k$ timesteps. Following that, the learnt representations are kept fixed and we fine tune the downstream policy learning algorithms on top of the fixed pre-trained representations. For the RL algorithm, as in [38] we use the TD3 + BC algorithm, since it has recently been shown to achieve state of the art performance on offline control tasks.

We provide additional results evaluating REPDIB on top of learnt representations in the visual pixel based offline RL setting. We implement REPDIB on top of the multi-step inverse dynamics objective [28], 1-step inverse dynamics [48] and the temporal contrastive learning based DRIML [39] objective. We show that REPDIB additionally compresses the learnt latent representations using the factorial bottlenecks, which makes the method quite effective and robust especially when there is additional exogenous information present in the observations [10]. We use different types of distractors in the offline datasets, where exogenous information can be either in the form of correlated background images or changing video distractors playing in the background during data collection. Figures 13 and 14 shows sample observations from the offline dataset.

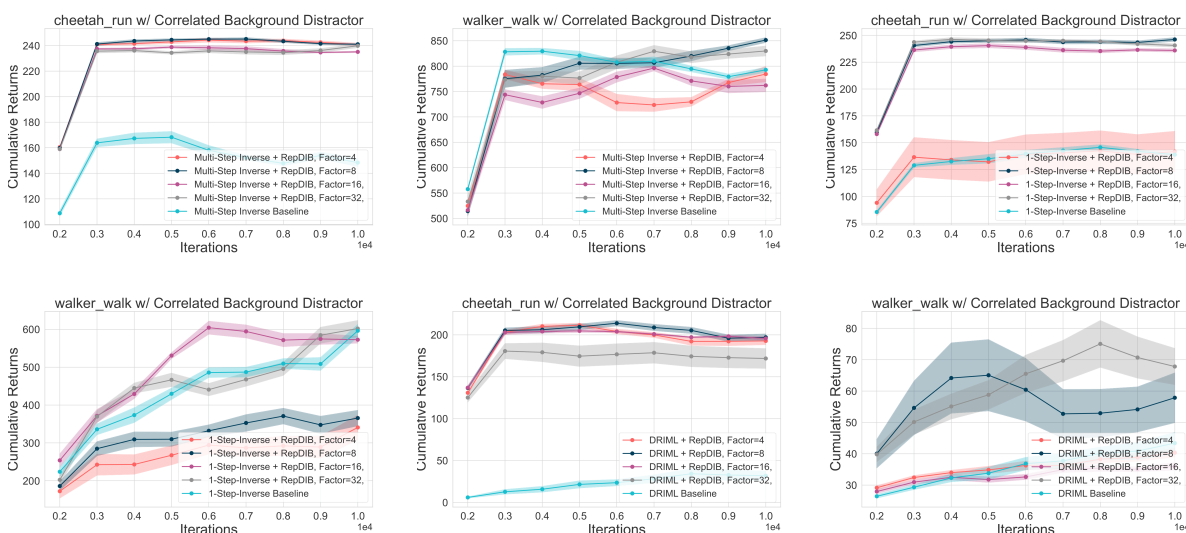


Figure 15: **Time correlated changing background image exogenous noise in offline datasets.** We consider a setting where the observations in the pixel based offline data consists of changing background image distractors in the background. The background exogenous noise is introduced during the data collection procedure. Such a setting requires learning of robust representations that can be invariant to the exogenous images. We consider 3 different representation learning objectives (a) Multi-Step Inverse (b) One-Step Inverse and (c) DRIML, where the encoders are pre-trained with these self-supervised objectives, followed by REPDIB. We show that for different factorial representations based on groups of factors 4, 8, 16, 32, the ability of these methods to learn robust representations due to REPDIB significantly increases, making them more robust to the exogenous offline datasets.

Experiment Results : Our experiments show that existing representation learning methods can suffer in presence of this exogenous noise being present, since the representations cannot fully avoid the distractors. In contrast, when adding REPDIB on top of the learnt representations, we find that compressed bottleneck representations can help in avoiding the distractors, improving the overall performance in the downstream offline RL tasks consisting of visual observations. We evaluate REPDIB on top of learnt representations using a 1-step inverse dynamics objective [48], a multi-step inverse dynamics objective [28, 10] and the temporal contrastive learning based DRIML objective [39]. Our results show that especially when exogenous noise is present in the observations, existing state of the art representation learning methods can suffer dramatically, leading to an overall degradation of performance. In contrast, addition of REPDIB can lead to improved performance due to bottlenecks that can capture factorial representations, while avoiding the exogenous distractors.

7 Significance of VIB and DIB for REPDIB

In this section, we include additional results based on ablation studies of the REPDIB objective. In figures 18 and 19 we include ablation studies where we compare REPDIB with *only* using the discrete information bottleneck (DIB) compared

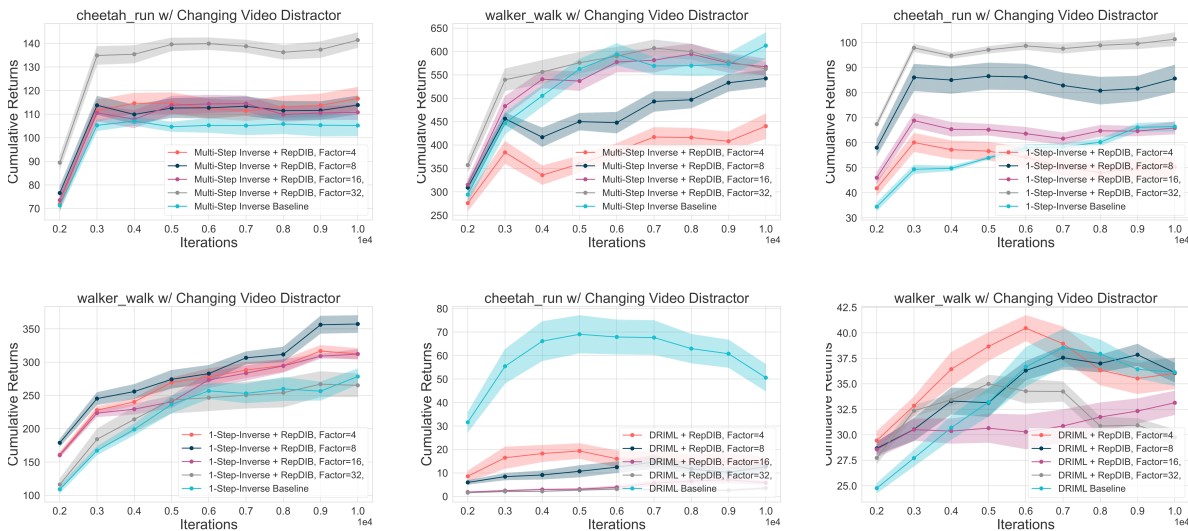


Figure 16: **Changing video exogenous noise in offline datasets.** We then consider a setting where there is background video distractors that changes per episode during data collection. Using REPDIB on top of the learnt representations from the same 3 different representation objectives, we find that in particular, the multi-step and one-step inverse models learns more robust representations compared to the DRIML objective. The changing background video distractors is considered to be a hard offline setting, since there is time correlated exogenous information continuously changing and playing in the background. We show that REPDIB improves the sample efficiency and overall performance of these methods, when used on visual offline data where learning robust representations plays a key role.

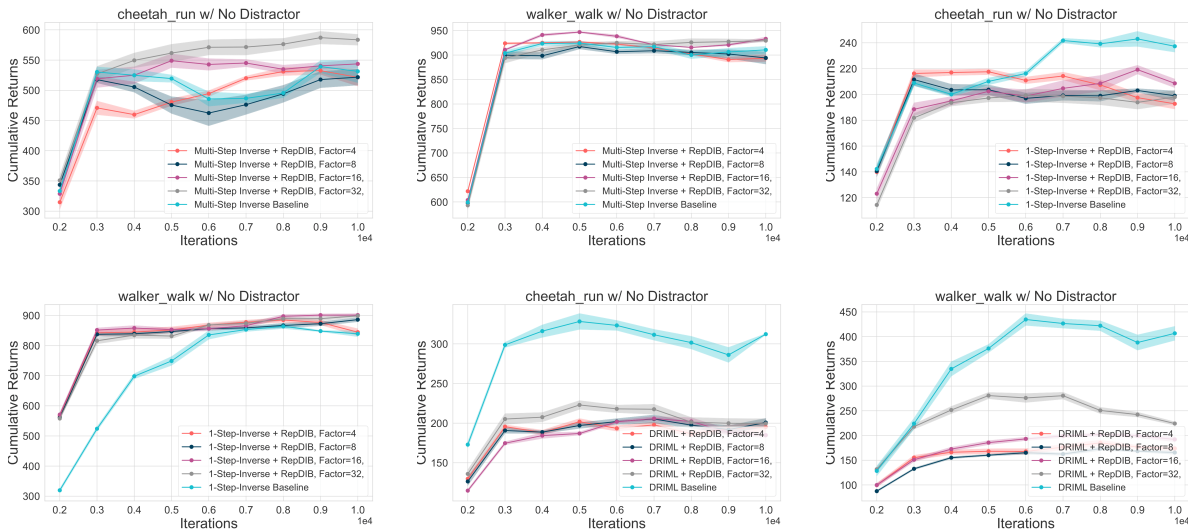


Figure 17: **Visual offline datasets from the v-d4rl benchmark [38] without any additional exogenous distractors.** We showed that REPDIB can learn effectively robust representations in presence of correlated exogenous noise as in figures 15 and 16. Here we show that without any distractors being present, REPDIB does not necessarily always outperform baselines, as shown in the results with the DRIML objective. This validates our claim that REPDIB based bottlenecks can be particularly effectively when observations consist of exogenous information, such that REPDIB can be used to learn more robustly. Without any distractors, it is not always necessary that REPDIB based representations would always outperform baselines without bottlenecks.

to *only* using the variational information bottleneck (VIB). We do this on top of several existing representation objectives as described in section 6.1. Experimental results show that the significance in performance improvement of REPDIB can primarily be achieved when we use the VIB bottleneck prior to the DIB bottleneck, as we have explained previously in the main draft. Without the combination of the two, simply using one of the bottlenecks does not lead to the expected performance improvements.

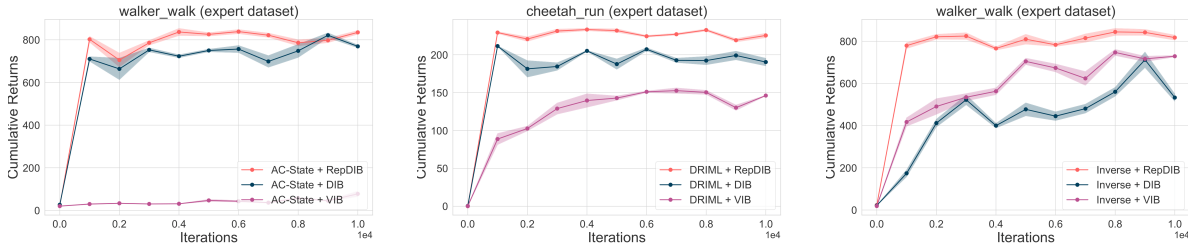


Figure 18: Ablation studies on the REPDIB bottleneck on time correlated exogenous distractors in the observations of offline datasets, as per the setup described in section 4.3

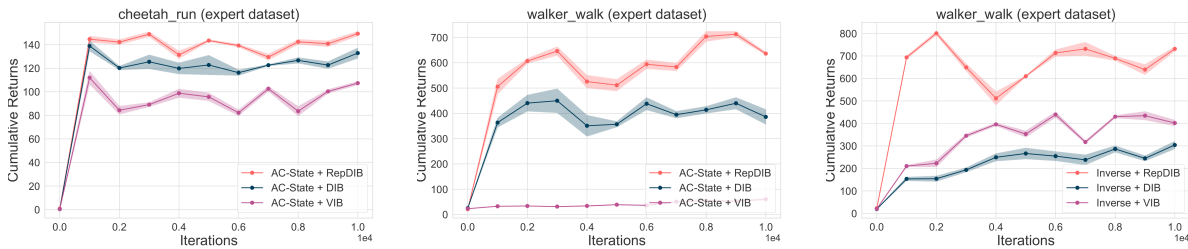


Figure 19: Ablation studies on the REPDIB bottleneck on changing background video based exogenous distractors in the observations of offline datasets, as per the setup described in section 4.3

7.1 Generalization on Continuous Control Tasks using URLB Benchmark

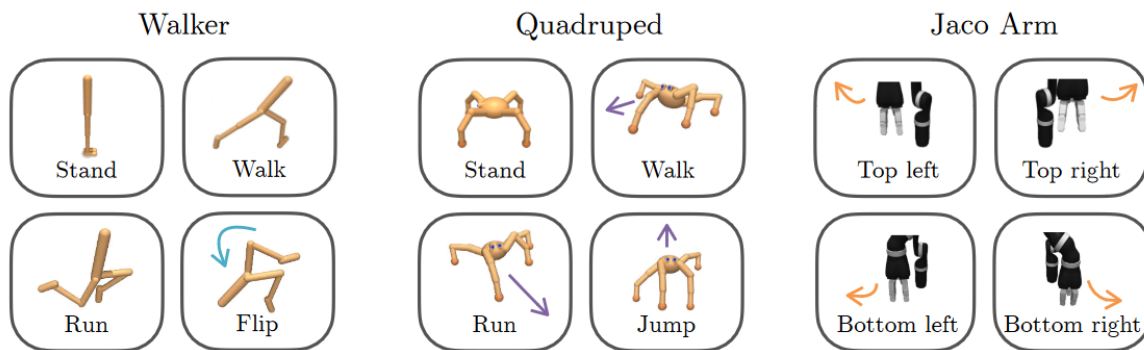


Figure 20: The three domains (walker, quadruped, jaco arm) and twelve downstream tasks.

Experiment Setup and Details : We follow the same set of domains and downstream tasks in [30] (See Figure 20). Specifically, from easiest to hardest, the domains and tasks are: **Walker** (*Stand, Walk, Flip, Run*): An improved planar walker based on the one introduced in [33]. In *Stand* task, the reward is a combination of terms encouraging an upright torso and some minimal torso height, and in *Walk* and *Run* tasks, the reward is proportional to forward velocity, while in *Flip* task, it is relative to angular velocity. **Quadruped** (*Stand, Walk, Jump, Run*): A quadruped within a 3D space. The reward function defined in quadruped is similar to that in walker, but quadruped is harder due to a high-dimensional state and action spaces and 3D environment. **Jaco Arm** (*Reach top left, Reach top right, Reach bottom left, Reach bottom right*): Jaco Arm is a 6-DOF robotic arm with a three-finger gripper that tests the ability to control the robot arm to perform

simple manipulation tasks. More detailed explanation refer to [30]. In Table 5 we present a set of hyper-parameters used in continuous control tasks.

Experiment Results : For the online continuous control tasks, we test for generalization using the URLB benchmark [30], on 12 different environments as shown in figure 20. In these experiments, representations are pre-trained on one environment for $100k$ pre-training steps, followed by fine-tuning both the RL algorithm and the encoder in a different environment. Existing baselines such as the ProtoRL [65] has already shown impressive performance compared to other baselines on the URLB benchmark. For more details on experiment setup and comparisons of ProtoRL with other baselines, see [30]. In this task, we take the open source code of the ProtoRL baseline and simply integrate REPDIB on top of the encoders, where we have different factors for learning representations. The goal of the experiments is to show that when using compressed representations that are structured and factorial in nature, the compression on the pre-training task helps in fine-tuning on other tasks when the same bottleneck is again applied. Figure 21 summarizes the ablation studies of REPDIB built on top of the ProtoRL baseline. Our results in figure 21 show that fine-tuning performance is mostly improved compared to the baseline, typically for higher factors of representation.

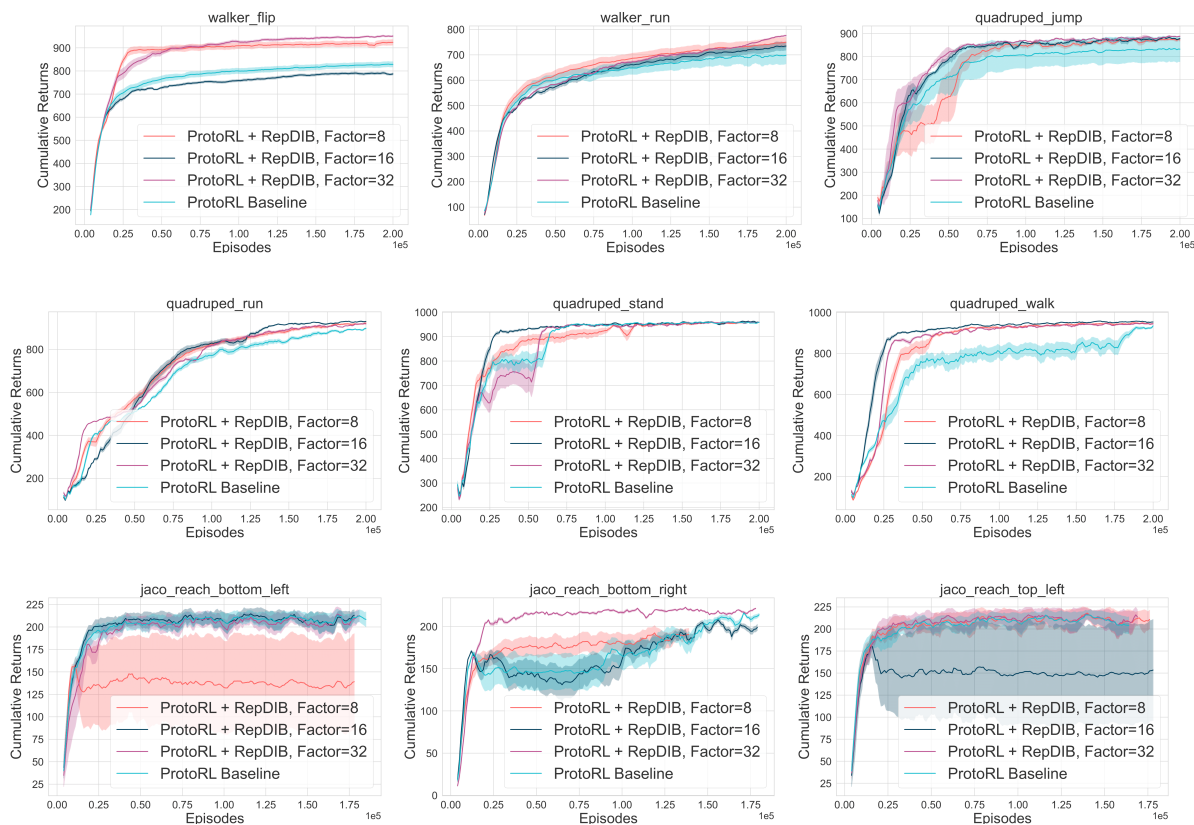


Figure 21: **URLB Benchmark for Continuous Control** Ablation analysis on the URLB benchmark, integrating REPDIB on top of the ProtoRL baseline with different factorizations of the discrete bottleneck. Our experiment results show that the factorization in representation, depending on the number of factors, can play a vital role in improving the performance on the generalization task.

7.2 Robot Arm Experiment

Robot Arm Experiment with Background Video Exogenous Distractors: The robot arm in our experiments moves in a grid with 9 different positions. We use two cameras to take images, for the dataset, one from the front side of the robot and the other with a top down view from above. We collect an image after each action is taken. The robot has 5 actions to take : move forward, backwards, right, left or stay in the current state. We use an episodic length of 500, ie, the robot arm moves for 500 steps after which we re-calibrate. The robot arm dataset is collected with a random uniform policy, for a total of 6 hours collecting 14000 samples.



Figure 22: Experiment setup for robot data collection in presence of exogenous or irrelevant background information. Figure shows three different images, with varying background information, for the robot arm, which are part of the collected dataset.

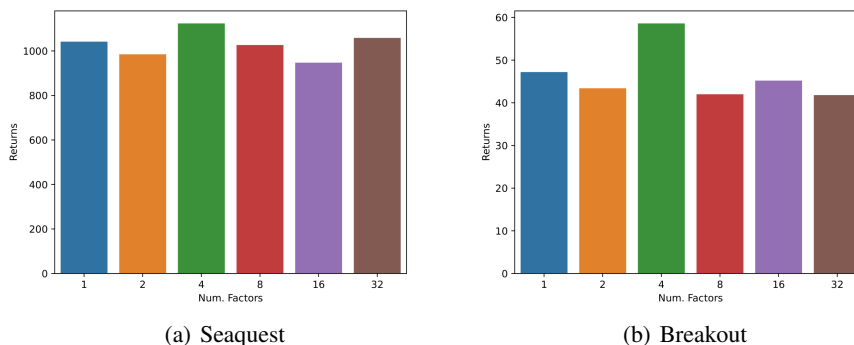


Figure 23: Here we show the effect of the number of discretization factors on the model performance for 2 different Atari games. On the ALE benchmark, we find that factor 4 usually outperforms other factors when learning representations with factorial structure.

For learning the representation ϕ given the images, we use a small convolutional neural network to get an estimate $\phi(x)$ of the images x . In addition to the CNN network, we further learned the latent state representation with a multi-step inverse dynamics model $p(a | \phi(x), \phi(x_k))$, which predicts actions, given current representation $\phi(x)$ and a future representation $\phi(x_k)$. The model is trained with a cross entropy loss, with the ground truth actions available in the dataset. We use a metric of classification accuracy for evaluating the performance of REPDIB.

7.3 Atari Benchmark with Exogenous Observations

Experiment Setup : We follow the experiment setup of decision transformers on the Atari domain following [7]. However, in addition to the environment observations from Atari games, we additionally augment the observations with an exogenous noise on the side. For this, we use CIFAR images placed on side of environment observations as exogenous noise. In Figure 24, we show example observations from atari games with exogenous noise added. The goal is to see the effect of REPDIB when integrated on top of a multi-step inverse dynamics objective for learning robust representations [28]. We keep most of the hyperparameter details same as used in [7]. They use episodes of fixed length during training - also referred to as the *context length*. We use a context length of 30 for Seaquest and Breakout. Similar to [7], we consider one observation to be a stack of 4 atari frames. To implement the multi-step inverse objective, we sample 8 different values for k and calculate the objective for each value of k , obtaining the final loss by taking the sum across all the sampled values of k . We do not feed the embedding for k in the MLP that predicts the action while computing the multi-step inverse objective.

Figure 23 shows the effect of different factors used in the discrete information bottleneck of REPDIB. We check the effect of the number of discretization factors on the model performance. The effect of the number of factors on overall performance of the Decision Transformer can vary depending on the game and domain.

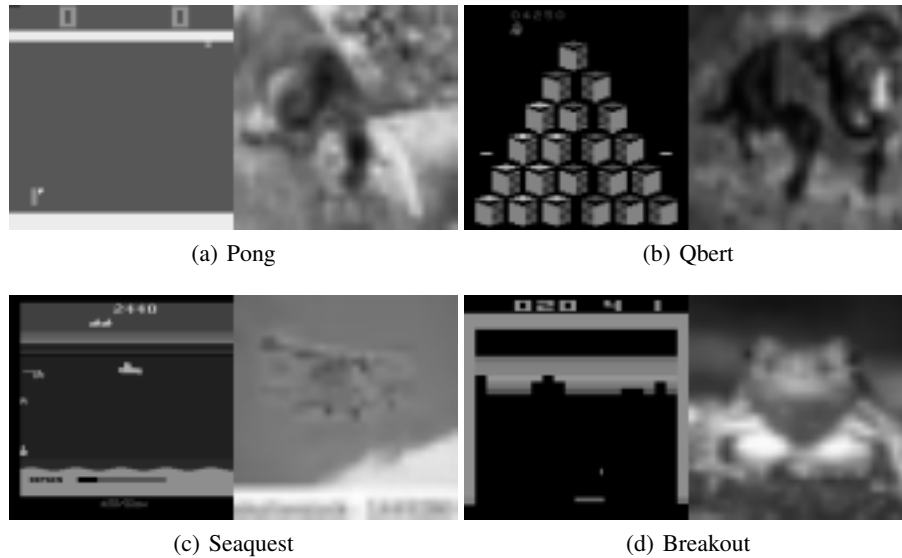


Figure 24: Example observations from 4 different Atari games, with exogenous images placed on the side of environment observations. We add exogenous noise to show the importance of learning robust representations using an information bottleneck following REPDIB.

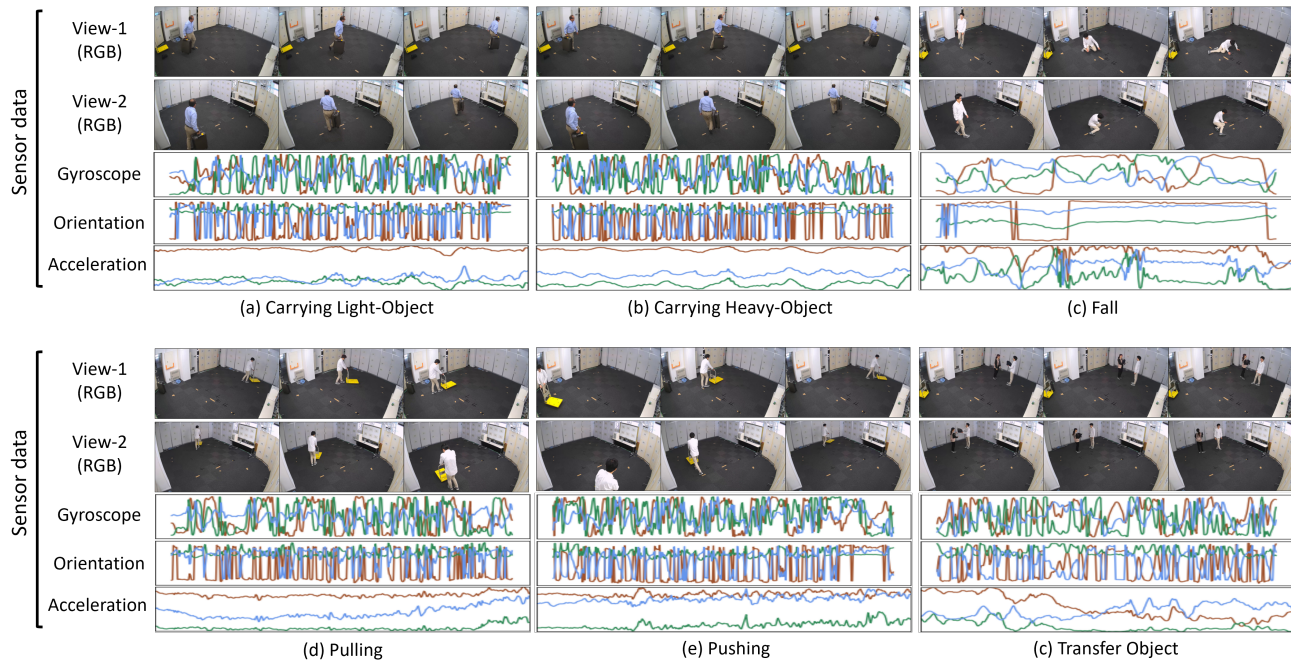


Figure 25: Sample data of human activities from MMAct dataset with five modalities: Visual views 1 & 2, Gyroscope, Orientation, and Acceleration.)

7.4 Multi-Modal Representation Learning on Human Activity Recognition Task

Dataset: MMAct dataset contains 37 activities (e.g., carrying objects, fall, kicking, talking on the phone, jumping, using PCs, sitting). Twenty people performed each activity five times, resulting in $37k$ data samples. All the activities are captured using data from seven modalities: four RGB views, acceleration, gyroscope, and orientation. We used data from two opposing RGB visual views, acceleration, gyroscope, and orientation modalities to train and test. MMAct dataset contains visually occluded data samples, which allows evaluating the effectiveness of HAR approaches for real-world settings. Human activity sample data are depicted in Figure 25.

Method	F1-Score (%)
SVM+HOG [44]	46.52
TSN (RGB) [62]	69.20
TSN (Optical-Flow) [62]	72.57
MMAD [25]	74.58
TSN (Fusion) [62]	77.09
MMAD (Fusion) [25]	78.82
Keyless [37]	81.11
HAMLET [18]	83.89
RepDIB+MM(Keyless)	71.35
RepDIB+MM(RepDIB+Uni)	84.96

Table 3: **Cross-session performance** comparison (F1-Score) of multimodal learning methods on MMAct dataset

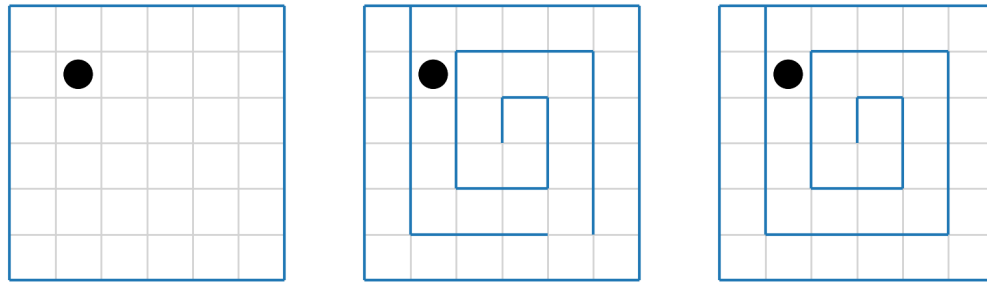
Experimental Setup for Multimodal Model Evaluation in Cross-Session Setting: In this supervised learning task, the model uses multimodal sensor data to recognize human activities. We extend state-of-the-art multimodal representation learning models to extract salient representation using RepDIB information bottleneck. We extended the baseline multimodal models in two ways to incorporate VQ bottleneck: **REP DIB+MM:** We extract multi-modal representations using existing models (e.g. Keyless [37] and HAMLET [18]) and then apply VQ bottleneck on the fused multi-modal representations. **REP DIB+MM(RepDIB+Uni):** We applied VQ bottleneck in two steps. First, we extract unimodal representations and apply VQ bottleneck to produce discretized unimodal representations. These discretized representations are fused and passed through a VQ bottleneck to produce task representations for the activity recognition. In the baselines, we used five modalities: two viewpoints of RGB videos and three wearable sensors (acceleration, gyroscope, and orientation). We evaluated all the baselines on the MMAct dataset in a cross-session evaluation setting and reported F1-Score of activity recognition task [25]. In the cross-session evaluation setting, the training and testing datasets can contain data from the same human subjects.

We train these models using cross-entropy loss. We use Adam optimizer with weight decay regularization and cosine annealing warm restarts learning scheduler, where the initial learning rate is set to $3e^{-4}$. To train the learning model on the MMAct dataset, we set the cycle length (T_0) and cycle multiplier (T_{mult}) to 30 and 2, respectively. We trained the models for 210 epochs in the distributed GPUs cluster environment, where each node contains 8 A100 GPUs. We used Pytorch and Pytorch-Lightning frameworks to implement all the models. To ensure reproducibility we a fixed seed.

Experimental Results: The results in Table 3 suggest that incorporating VQ bottleneck on the existing multi-modal learning model (Keyless) degrades the F1-score of activity recognition from 81.11% to 71.35. However, applying the VQ bottleneck both on the unimodal and multimodal representations improves the performance of the models compared to the models that do not use REP DIB or use REP DIB only on the multimodal representations. For example, REP DIB + MM(RepDIB + Uni) model uses the same HAMLET model and applies REP DIB on the unimodal and multimodal representations. REP DIB + MM(RepDIB + Uni) improves the F1-score of activity recognition to 84.96 and outperforms all the evaluated multimodal models. Thus, hierarchical VQ bottlenecks can help to extract salient multimodal representation for accurately recognizing activities.

7.5 Maze Navigation Tasks

Maze Navigation Tasks We develop three kinds of environments for maze navigation tasks: *GridWorld*, *SpiralWorld*, *LoopWorld* (see Figure 26). All of these environments share the same action space and state space, but their dynamics are slightly different. *Gridworld* is the easiest task that without any walls so that the agent can go wherever it wants. *SpiralWorld* is the hardest one that has spiral-shaped walls blocking the path of the agent, where the agent can only navigate along the spiral grid. *LoopWorld* is a variant of *SpiralWorld* in which the agent can pass through a vacancy in the bottom right corner of the spiral-shaped wall. The task is to choose from one of four directions to travel in at each timestep. The reward function given is -1 at all steps until it reaches the goal where it receives a reward of 0 and the episode is terminated. During pre-training stage, we learn the state representations on *GridWorld* with the data collected by a random policy. During the fine-tuning stage, the agent is trained to reach a goal from a small finite set of training goals, and the agent is tasked with reaching a fixed goal at the center of the maze during evaluation. In Table 4 we present a set of hyper-parameters used in maze navigation tasks.

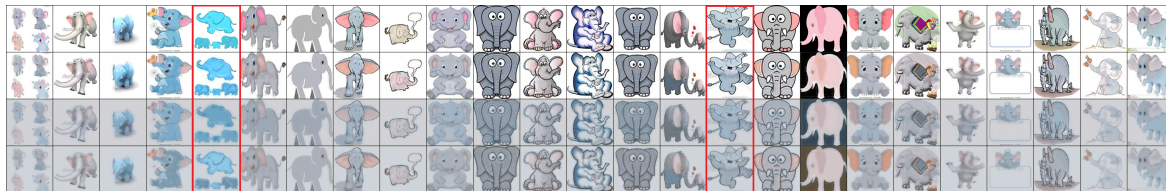


(a) GridWorld

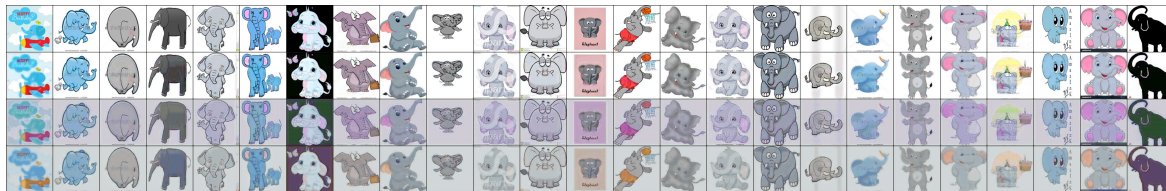
(b) LoopWorld

(c) SpiralWorld

Figure 26: Three environments in maze navigation tasks. The blue lines represent the walls.



(a) Brightness and Details



(b) Different colors

Figure 27: PACS-cartoon-elephant dataset example to demonstrate factorized representations. Top row: Original Image; Second row: Reconstructed Image without substitution; Third row: Reconstructed Image with one groups of discrete codes substituted by zero vectors; Last row: Reconstructed Image with the other groups of discrete codes substituted by zero vectors.

8 Demonstrating Factorial Representation

We demonstrate that with the discrete factorial information bottleneck, the agent is capable of learning factorial representations on real world data. We provide more details as follow.

Experiment details To investigate whether the agent has the ability to learn semantic factorial representation with REPDIB, we use the cartoon domain images from a benchmark dataset called PACS [31], where only the elephant category is utilized for training and evaluation for the purpose of the intuitively illustration. The pixel-based input, with the size of 224×224 , is first passed through an encoder (consists of CNN layers with the resnet block) to obtain its latent representation with the dimension of 32, then latent representation is quantized into two groups of discrete codes, where the codebook size is 512. After that, two groups of discrete codes are concatenated to obtain the representation, and finally passed through a decoder network (consists of CNN layers with the resnet block). Here we used reconstruction loss (MSE loss) combining with the loss for vector quantization to train the network. For visualizing the semantic meaning of different groups, we randomly sample 25 pictures from the dataset, and pass the images into the network to obtain reconstruction of the images. Ideally, we would like to know whether different groups capture different semantic meaning of one image. For this purpose, we used zero vector to substitute one group of the discrete codes and acquire the reconstructed image by concatenating it with the other group of the discrete codes. As a consequence, we have three reconstructed images in total,

as shown in Figure 27.

Experiment Results Figure 27 shows the reconstructed images from a trained decoder operating on a discretized 2-factor representation. We find that different factors capture different semantic information. As an example, it is obvious to see that there are 4 elephants in the fifth column in Figure 27(a), where the elephant at the top and the elephant at the bottom-middle are brighter than the other two elephants. For this image input, factor 1 tends to only capture the shape of the elephant without the brightness, while factor 2 capture specific details of each elephant. The similar observation can be found in the 16th column, where factor 1 captures the “shadow” in the picture, and factor 2 captures the brightness of elephant’s skin. Another example is in Figure 27(b), it is shown that two factors learn “green” and “purple” separately for reconstructing “black”, and two factors learn “pink” and “orange” separately for reconstructing “red”.

9 Explanation and Significance of REPDIB

We would like to provide further clarification about the significance of our work. In this work, we do not propose any new representation learning objective; rather we simply propose that discrete information based bottlenecks can be significant when it comes to learning representations. Moreover, an approach based on REPDIB is demonstrated to be even more impactful especially when the learnt representation needs to discard exogenous or irrelevant information from the observations. We demonstrate this across a range of experiments, not only based on RL, but also based on other tasks such as human activity recognition. Our experiments however are primarily based on RL benchmarks, where we demonstrate that REPDIB can be easily applied on top of any learnt representations. To do this, we take existing baseline approaches proposing representation learning objectives and demonstrate the ease with which REPDIB can be integrated on top of learnt representations.

We emphasize that although information bottlenecks has been studied extensively in past literature, the use of discrete information bottleneck is rather new; and moreover to apply bottlenecks on top of representation learning objectives, especially to discard exogenous information, has been little studied in the past. Our aim is to propose information bottleneck, which not only captures factorial or compositional representations, but also plays key role in extracting only the relevant latent representation; and most importantly, can be suitably applied on any deep RL algorithm relying on additional representation learning module.

10 Visualization

In our experiments, we constantly find that the use of a variational information bottleneck (VIB) prior to the discretization bottleneck significantly helps performance of REPDIB

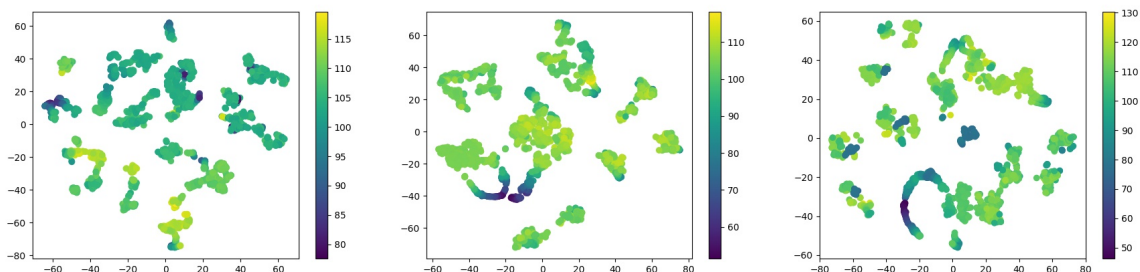


Figure 28: **Comparing Visualizations with and without bottleneck representations** t-SNE of latent spaces in the JacoReachTopRight task learned with Proto-RL (left t-SNE), REPDIB (middle t-SNE), and REPDIB with VIB (right t-SNE) after training has completed, color-coded with predicted state values (higher value yellow, lower value purple).

11 Hyperparameter Details

Hyper-parameter	Value
Size of Maze	6×6
Mini-batch size	128
Discount (γ)	0.99
Optimizer	Adam
Learning rate	3×10^{-3}
Critic target EMA rate (τ_Q)	0.01
Features dim.	128
Hidden dim.	128
Number pre-training frames	1×10^4
Number of discrete codes	50
Number of groups	8, 16, 32
VIB coefficient	0.01

Table 4: A set of hyper-parameters used in maze navigation tasks.

Common hyper-parameter	Value
Replay buffer capacity	10^6
Seed frames	4000
n -step returns	3
Mini-batch size	1024
Seed frames	4000
Discount (γ)	0.99
Optimizer	Adam
Learning rate	10^{-4}
Agent update frequency	2
Critic target EMA rate (τ_Q)	0.01
Features dim.	1024
Hidden dim.	1024
Exploration stddev clip	0.3
Exploration stddev value	0.2
Number pre-training frames	up to 2×10^6
Number fine-tuning frames	up to 2×10^6
Number of discrete codes	50
Number of groups	8, 16, 32
VIB coefficient	0.01

Table 5: A set of hyper-parameters used in continuous control tasks.

Common hyper-parameter	Value
n -step returns	3
Mini-batch size	256
Seed frames	4000
Discount (γ)	0.99
Optimizer	Adam
Learning rate	3×10^{-4}
Critic target EMA rate (τ_Q)	0.01
Features dim.	256
Hidden dim.	1024
Number pre-training frames	1×10^5
Number fine-tuning frames	1×10^5
Number of discrete codes	512
Number of groups	4, 8, 16, 32
VIB coefficient	0.01

Table 6: A set of hyper-parameters used in offline tasks.