Value Implicit Pretraining does not learn representations suitable for Reinforcement Learning

Harshit Sikchi[*]
The University of Texas at Austin
hsikchi@utexas.edu

Siddhant Agarwal[*]
The University of Texas at Austin
siddhant@cs.utexas.edu

Peter Stone
The University of Texas at Austin
Sony AI
pstone@cs.utexas.edu

Scott Neikum
University of Massachusetts Amherst
sniekum@cs.umass.edu

Amy Zhang
The University of Texas at Austin
amy.zhang@austin.utexas.edu

Abstract

Leveraging large-scale internet data to bootstrap representations for RL presents a reasonable path forward to learning general-purpose agents. Unlike computer vision and NLP, the sequential nature of the RL problem makes it unclear what the learning rule should be. Goal-Conditioned RL (GCRL) presents a self-supervised objective for RL that can allow agents to learn representation suitable for planning to arbitrary states. Traditional GCRL objectives require action labels which are usually missing from sequential data on internet (eg. videos). A recent approach, Value Implicit Pretraining (VIP), presents a new objective to learn optimal value functions without requiring action labels. However, our findings in this paper suggest that VIP fails to learn the correct representations in simple domains. This finding motivates us to conduct a detailed investigation, and through formal arguments we establish why VIP shows this anomaly. We propose a simple yet effect alternative, DERAIL, that indeed learns optimal value functions and subsequently representations suitable for RL.

1 Introduction

Fields like Computer Vision and Natural Language Processing have immensely benefited from utilizing large-scale vision and language data. Unfortunately, such developments in reinforcement learning and robotics have been limited. Recent works like Walke et al. (2023); Padalkar et al. (2023) have made significant strides in creating large-scale datasets that can be used for robotics. To learn representations from such large-scale data that can be transferred easily to different domains and downstream tasks requires the representations to be independent of underlying action space i.e. learning representations only from observations. R3M (Nair et al., 2022), VIP (Ma et al., 2022), LIV (Ma et al., 2023), ICVF (Ghosh et al., 2023) are a few successful works that use observations from large-scale egocentric video data to learn value functions parameterized using encoders. These encoders can then be transferred to other domains and can be used either to define reward functions or as a backbone encoder for behavior cloning.

[*] Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Learning representations for RL requires navigating a number of challenges: (1) What is the right representation learning objective? (2) What are we going to use the representation for? (3) How do we evaluate those representations. Unfortunately, these things are murkier in the field of reinforcement learning when compared to supervised learning. A promising approach to this problem is to use Goal-Conditioned RL Kaelbling (1993) as an self-supervised objective for representation learning. The idea is to learn compact representations that can allow for planning from any state in the environment to any other state — effectively compressing the observations into representations suitable for planning. Learning representations through GCRL using offline datasets that lack actions is challenging as most objectives require actions to be known (Eysenbach et al., 2021; Sikchi et al., 2023a). A recently developed line of work, Value Implicit Pretraining (VIP) (Ma et al., 2022) leverages Fenchel-Rockefeller duality by treating RL as a convex program and derives an action-free objective for GCRL that is later adapted to representation learning.

In our work, we demonstrate that representations learnt through Value Implicit Pretraining are lacking and unsuitable for planning. VIP's anomaly arises as a result of imposing an information bottleneck in the primal objective requiring assumptions that don't hold in practice, to convert to a dual objective suitable for representation learning. We hypothesize that the perceived success of VIP is due to using expert-like trajectories in its offline dataset of transitions and learning representation using a time-contrastive objective between neighbouring observations. Our empirical experiments confirm our findings that the representation learned through VIP indeed fail on simple low-dimensional tasks. Next, gathering the insights from our argument, we develop a simple, principled, and action-free objective for representation learning that we then use to overcome the limitations of VIP. Our proposed objective is motivated by the recently proposed dual perspective of reinforcement learning. We show that improves representation learning from a series of qualitative and quantitative experiments.

2 Related Works

Representation Learning in RL can be broadly classified into two categories: (a) offline pretraining (Ma et al., 2022; Nair et al., 2022) and (b) using an auxilliary loss (Schwarzer et al., 2021; Agarwal et al., 2021; Agarwal et al.) over the RL loss. The goal of both these paradigms is to induce an inductive bias on the representation space by using reward-free interaction data. There are a variety of auxilliary objectives that can be added to produce desired properties in the representations majority of them being contrastive objectives (Schwarzer et al., 2021; Agarwal et al.; Srinivas et al., 2020) aiming for sample efficiency (Schwarzer et al., 2021; Srinivas et al., 2020), generalization (Agarwal et al., 2021; Agarwal et al.) and temporal consistency (Zhao et al., 2023). While these methods do introduce some interesting properties in the representation space and observes gains in sample efficieny and generalization, these do not look into pre-training task-agnostic generalizable encoders from offline data.

With the availability of large scale datasets like Ego4D (Grauman et al., 2022) and Epic Kitchens (Damen et al., 2018), several methods have been developed that look learn representations from large-scale pretraining trying to bridge the gap between RL and fields like computer vision and NLP. RRL (Shah & Kumar, 2021) and VC1 (Majumdar et al., 2024) are some methods that have attempted using classical computer vision techniques for representation learning. However, these works do not take into account the sequential nature and temporal data. R3M (Nair et al., 2022) used Ego4D to learn representations using Time Constrastive Loss on the trajectories of Ego4D while VIP (Ma et al., 2022) and the follow up work LIV (Ma et al., 2023) introduced this temporaral consistency implicitly by learning a goal-conditioned value function for the trajectories in the large-scale dataset.

3 Preliminaries

We consider a learning agent in a Markov Decision Process (MDP) (Puterman, 2014; Sutton & Barto, 2018) which is defined as a tuple: $\mathcal{M} = (\mathcal{O}, \S, , P, R, \gamma, d_0)$ where $\S$ and denote the state and action spaces respectively, P denotes the transition function with $P(s'|s,a)$ indicating the probability of transitioning from s to s' taking action a; R denotes the reward function and $\gamma \in (0,1)$ specifies the discount factor. We use $o \in \mathcal{O}$ to denote the space of observations, where an observation is generated as a stochastic function of underlying state. The reinforcement learning objective is to obtain a policy $\pi : \mathcal{O} \to \Delta()$ that maximizes expected return: $\pi \sum_{t=0}^{\infty} \gamma^t r(o_t, a_t)$, where we use $\mathbb{E}_\pi$ to denote the expectation under the distribution induced by $a_t \sim \pi(\cdot|o_t), o_{t+1} \sim p(\cdot|o_t, a_t)$ and $\Delta()$ denotes a probability simplex supported over . W

Dual formulation of RL Dual RL (Sikchi et al., 2023b), also called Distribution Correction Estimation (DICE) (Nachum & Dai, 2020) present a family of principled off-policy algorithms that can leverage data from arbitrary sources to learn optimal policy. Dual RL works by first considering the following two convex program formulations of regularized reinforcement learning in the form of `primal-Q` and `primal-V`:

$$\begin{aligned}&\max_{\pi} J(\pi) = \max_{\pi} \big[\max_{d}\ \mathbb{E}_{d(o,a)}[r(o,a)] - \alpha d(o,a) d^O(o,a) \\ &\text{s.t } d(o,a) = (1-\gamma) d_0(o).\pi(a|o) + \gamma \textstyle\sum_{s',a'} d(o',a') p(o|o',a') \pi(a|o),\ \forall o \in \mathcal{O}, a \in \big].\end{aligned} \tag{1}$$

and,

$$\begin{aligned}&\max_{d \geq 0}\ \mathbb{E}_{d(o,a)}[r(o,a)] - \alpha d(o,a) d^O(o,a) \\ &\text{s.t } \textstyle\sum_{a \in \mathcal{A}} d(o,a) = (1-\gamma) d_0(o) + \gamma \sum_{(o',a') \in \mathcal{O} \times} d(o',a') p(o|o',a'),\ \forall o \in \mathcal{O}.\end{aligned} \tag{2}$$

The constraints above represent the Bellman flow conditions that any valid visitation distribution should satisfy. That is, the visitation distribution should be induced by some policy under the dynamics of the environment. Applying Lagrangian duality and using convex conjugates result in respective unconstrained optimization problems for solving regularized RL:

$$\max_{\pi} \min_{Q} (1-\gamma) o \sim d_0, a \sim \pi(o) Q(o,a) + \alpha (o,a) \sim d^O f^* \left([Q(o,a) - Q(o,a)] / \alpha \right), \tag{3}$$

and,

$$\min_{V} (1-\gamma) o \sim d_0 V(o) + \alpha (o,a) \sim d^O f_p^* \left([\mathcal{T} V(o,a) - V(o))] / \alpha \right), \tag{4}$$

where denotes Bellman operator with policy π and reward function r such that $Q(o,a) = r(o,a) + \gamma o' \sim p(\cdot|o,a), a' \sim \pi(\cdot|o') Q(o',a')$ and $\mathcal{T} V(o,a) = r(o,a) + \gamma s' \sim p(\cdot|s,a) V(o')$.

4 Relating Value Implicit Pretraining to Optimal Value Function Learning

In this section, elucidate why a classical implementation of VIP objective does not quite learn optimal value functions. Our analysis below identifies why VIP value functions perform worse and later propose guidance on how to train representations with a dual objective.

4.1 What objective is VIP trying to learn?

VIP learns goal-conditioned value functions by leveraging the GCRL as a convex optimization problem. Following the DICE framework, it constructs a convex objective by regularizing the distribution matching problem with linear constraints on visitation distribution to generate a corresponding dual objective which is action-free. Additionally, VIP introduces an information bottleneck on observations, thus maximizing the expected return of reaching various goals in the environment.

$$\begin{aligned}\max_{d,\phi} \quad & \mathbb{E}_d[r(o;g)] - D_{KL}(d(\phi(o), a; \phi(g)) \| d^D(\phi(o), a; \phi(g))) \\ s.t \quad & \sum_{a} d(\phi(o), a; \phi(g)) = (1-\gamma)\mu_0(o,g) + \gamma \sum_{o',a'} T(o|o',a') d(\phi(o'), a'|\phi(g))\end{aligned} \tag{5}$$

The idea is that ϕ encodes sufficient statistics about observations and goals in such a way that still allows it to solve the GCRL problem as well as possible. We ask the question *if VIP indeed succeeds in maximizing this objective below.*

4.2 Issues with VIP under the representation learning objective

To identify the issues with the representation learning framework presented by VIP, we shall go through the broad derivation of the representation learning objective from the dual objective. The

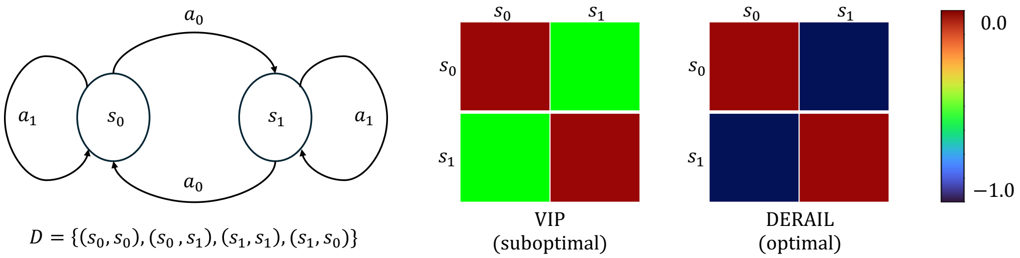

Figure 1: (left) A simple 2-state 2-action MDP with a reward, and the dataset containing all 2 state trajectories obtained from the MDP. (right) A comparison of the goal conditioned value functions $V(s_1, s_2)$ learnt by VIP and DERAIL for the goal conditioned reward, $r(s, g) = 0$ if $s = g$ and -1 otherwise. *VIP produces suboptimal value functions while DERAIL learns the optimal value function.*

Lagrangian dual of the above primal representation learning objective with some algebraic rearrangement is shown below. Although VIP uses KL divergence, we will be presenting the derivation for any general f-Divergence.

$$\min_{V} \max_{\phi, d \geq 0} (1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V(\phi(o), \phi(g))] + \mathbb{E}_{d(\phi(o),a;\phi(g))}[r(o;g) + \gamma V(\phi(o'), \phi(g')) \\ -V(\phi(o), \phi(g))] - D_f(d(\phi(o), a; \phi(g)) || d^D(\phi(o), a; \phi(g)) \quad (6)$$

The inner maximization w.r.t d assumes a closed-form solution, denoted by f^*_{VIP} which reduces the representation learning objective to a two-player game between V and ϕ.

$$\max_{\phi} \min_{V} (1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V(\phi(o), \phi(g))] + \mathbb{E}_{d(\phi(o),a;\phi(g))}[f^*_{VIP}(r(o;g) + \gamma V(\phi(o'), \phi(g')) \\ -V(\phi(o), \phi(g)))] \quad (7)$$

To ease learning, it is assumed that the optimal value function for any ϕ has the following structure $V^*(o, g) = -\|\phi(s) - \phi(g)\|$.

$$\max_{\phi} (1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V^*(\phi(o), \phi(g))] + \mathbb{E}_{d(\phi(o),a;\phi(g))}[f^*_{VIP}(r(o;g) + \gamma V^*(\phi(o'), \phi(g')) \\ -V^*(\phi(o), \phi(g)))] \quad (8)$$

We use a simple symmetric MDP example below with two states where one state transitions to another with a particular action or remains at the same location with the other action. The use of symmetric MDP is to ensure that the optimal value functions are representable by the structural assumption of VIP. We will use this MDP as a proof by counterexample to demonstrate VIP's failure to learn optimal value functions.

Issue 1: VIP ignores the positivity constraint $d \geq 0$ changes the fixed point of optimization

The inner maximization with respect to the visitation distribution d admits a closed-form solution. This closed-form solution can differ significantly if the positivity constraint of $d \geq 0$ is ignored. We compare the functional forms of the conjugate function used in VIP vs the true conjugate function under positivity constraints in Table 1.

Divergence Name	Generator $f(x)$	Conjugate in VIP $f^*_{VIP}(y)$	True Conjugate $f^*_p(y)$
Reverse KL	$x \log x$	$\log xe^{y-1}$	$e^{(y-1)}$
Squared Hellinger	$(\sqrt{x} - 1)^2$	–	$\frac{y}{1-y}$
Pearson χ^2	$(x-1)^2$	$\frac{(y+1)^2}{2}$	$\max(\frac{y}{2} + 1, 0)y - (\max(\frac{y}{2} + 1, 0) - 1)^2$

Table 1: VIP's use of conjugate functions vs the true conjugate under positivity constraints. '–' denotes divergences not discussed in VIP.

Issue 2: Assumes a structure on the optimal value function that will be true regardless of the representations ϕ to simplify optimization

In two-player game or bilevel optimization where one variable depends on the value of another variable (in our case $V(\phi(o), \phi(g))$ is a function of phi), assuming an analytical relation on the fixed point (eq. $V^*(\phi(o), \phi(g)) = -\|\phi(o) - \phi(g)\|$) can lead to substantially different fixed point solution. Figure 1 shows the value function learned by VIP and value function learned by the method we propose later in this work. VIP converges to an incorrect value function.

In general, without imposing structural assumption on the value function, VIP objective remains a two-player game (Eq 7) bringing optimization challenges. Furthermore, the sampling distribution required in VIP is in the space of encoded representations and contributes to gradient updates. The traditional implementation ignores this by assuming no conflicting encoded observations (no embedding collisions i.e $\phi(o) \neq \phi(o') \forall o, o' \in \mathcal{O}$). This is additionally also a requirement when the dynamics are stochastic as the constraints in Eq 5 use transitions defined in the space of unencoded observations. This effectively means that if there were a simple two state MDP as in Figure 1 with D observations of each state, VIP would necessitate compression to $\log(2D)$ bits instead of the sufficient 1-bit representation of the MDP, thus losing the compression capability afforded by the structure of MDP. Our work aims to get away with these assumptions and propose a representation learning objective that retains the benefits of a single-player learning objective.

Algorithm 1 DERAIL

```
Init $V_\phi(s, g)$, conservatism $\lambda$
Let $\mathcal{D} = \hat{\rho} = \{(s, a, s')\}$ be an offline dataset.
for $t = 1..T$ iterations do
    Train $V_\phi$ via Orthogonal gradient update on Eq. 13
end for
return $\phi$
```

5 DERAIL: Learning optimal value functions with Dual-V Learning

Understanding the limitations of VIP, motivated by the dual framework, we turn to presenting a simple action-free objective for representation learning. Our key insight that the issues of learning optimal value function can be mitigated by first deriving a mathematically sound dual objective and then imposing a information bottleneck as opposed to imposing an information bottleneck on the primal and using approximation to derive the dual objective.

We consider the dual-RL objective Sikchi et al. (2023b) reformulated for goal-conditioned RL:

$$\max_d \mathbb{E}_d[r(o; g)] - D_f(d(o, a; g) \| d^D(o, a; g)) \tag{9}$$

where d represents the visitation distribution and is subject to the traditional bellman flow constraints that ensure d is induced by some policy respecting dynamics of the environment. Here the constraints become:

$$\sum_a d(o, a; g) = (1 - \gamma)\mu_0(o, g) + \gamma \sum_{o', a'} T(o|o', a') d(o', a'|g) \tag{10}$$

Computing the dual of the problem is easy, and can be done following the same steps as in Dual-RL (Sikchi et al., 2023b). Using straightforward algebraic manipulations we can rewrite the above equation as:

$$\min_V \max_{d \geq 0} (1 - \gamma)\mathbb{E}_{\mu_0(o;g)}[V(o, g)] + \mathbb{E}_{d(o,a;g)}[r(o; g) + \gamma \sum_{o'} T(o'|o, a, g)V(o', g) - V(o, g)] \\ - D_f(d(o, a; g||d^D(o, a; g) \tag{11}$$

The distributions in the above equation are all conditioned on a particular goal. The inner maximization problem w.r.t. d has a analytical solution and leads us to the final optimization objective:

$$\min_V (1 - \gamma)\mathbb{E}_{\mu_0(o;g)}[V(o, g)] + \mathbb{E}_{d(o,a;g)}[f_p^*(r(o; g) + \gamma \sum_{o'} T(o'|o, a, g)V(o', g) - V(o, g))] \tag{12}$$

With the action-free GCRL objective in hand, we now directly impose an information bottleneck on the observations by enforcing value function predictions to only use the encoded representations ($\phi(o)$) of observations.

Lemma 5.1. *Value function learning with the DERAIL objective converges to optimal value function under sufficient representation capacity of the information bottleneck.*

Proof. The derivation follows from leveraging the strong duality argument from Section B.1.4 in Sikchi et al. (2023b) along with the assumption of lemma that sufficient representation capability allows representing all value functions.

The objective for representation learning using observational data, DERAIL, can be written as:

$$\min_{\phi}(1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V_\phi(o,g)]+\mathbb{E}_{d(o,a;g)}[f_p^*(r(o;g)+\gamma\sum_{o'}T(o'|o,a,g)V_\phi(o',g)-V_\phi(o,g))] \quad (13)$$

Interestingly, the objective we obtain indicates an almost opposite behavior than the VIP objective which maximizes a monotonic function of bellman error, whereas DERAIL objective minimizes it.

5.1 Practical Algorithm

Learning representations from offline datasets require tuning conservatism akin to offline RL algorithms. Following Sikchi et al. (2023b), we incorporate conservatism by a linear weighting between the two terms in the objective:

$$\min_{\phi}(1-\lambda)(1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V_\phi(o,g)]+\lambda\mathbb{E}_{d(o,a;g)}[f_p^*(r(s;g)+\gamma\sum_{s'}T(o'|o,a,g)V_\phi(o',g)-V_\phi(o,g))] \quad (14)$$

We instantiate our algorithm (Algorithm 2) using the Pearson χ^2 divergence for which the f_p^* takes the following closed form:

$$f_p^*(y) = \max\left(\frac{x}{2}+1,0\right)x - \left(\max\left(\frac{x}{2}+1,0\right)-1\right)^2 \quad (15)$$

Substituting the above form of f_p^* in Eq. 13 gets us the practical objective we use in this work. To optimize Eq 13 we use orthogonal gradient updates that have been shown to be more effective in practice in finding the fixed point of the objective Mao et al. (2024) compared to semi-gradient updates Sikchi et al. (2023b). Prior works have found feature co-adaptation between features of the current state and the next state, leading to gradients of $V(o',g)$ and $V(o,g)$ canceling out. Orthogonal gradient updates fix this by considering the projection of the gradient of the next observation in the orthogonal direction to the gradient of the current observation.

We parameterize value functions by considering two representation bottlenecks in this work: (a) L2/Eucleadian ($V_\phi(o,g) = -\|\phi(o)-\phi(g)\|$) used in Ma et al. (2022), and (b) Multilinear used in Ghosh et al. (2023) ($V_\phi(s,g) = \phi_1(s)\phi_2(g)\phi_3(g)$) where $\phi_1(s) \in \mathbb{R}^d$,$\phi_2(g) \in \mathbb{R}^{d\times d}$ and $\phi_3(g) \in \mathbb{R}^d$. For multilinear representations we use ϕ_1 as the resulting representation encoder. The Eucleadian bottleneck linearizes the value function in representation of observation but has the downside of enforcing symmetric value functions $V(o,g) = V(g,o)$, a condition often violated in practice. For this reason, we consider Multilinear representation that allow for assymetric value function learning while still imposing an information bottleneck.

6 Experiments

Our experiments aim to validate the arguments made in the paper about the failure of VIP in learning correct representations for reinforcement learning and demonstrate the effectiveness of the proposed method DERAIL. To this end, we consider a number of MuJoCo tasks with freely accessible datasets Fu et al. (2020). We do not use the Ego4D dataset used in as the version of the dataset used in the paper is not available to public. Our results are not limited by the consideration of simulated tasks, as any representation learning method should learn meaningful representation invariant of domain being used. Indeed, recent work () shows that representations learned on states (already a compact representation) can speed up RL. Simulated tasks gives us the ability to do a more detailed analysis of OOD generalization capabilities of the learned representation. Our experiments below evaluate the learned representations from VIP and `DERAIL` both qualitatively and quantitatively.

Datasets: We use the D4RL datasets and consider the problem of learning representations that allow the agent to plan from any state in the dataset to any other state. Our preliminary results investigate representations learned on the following datasets — halfcheetah-medium-expert, hopper-medium-expert, walker-medium-expert, ant-medium-expert.

6.1 Nature of Learned Representations

In this section, we investigate whether the representations learned by are meaningful. To do so, we train each method to convergence for 100k gradient updates on all the datasets and use the encoder (the information bottleneck of the value function) to generate 16 and 32 dimensional representation of states in the environment. For the multilinear representation we discard ψ which encodes the task information of which goal to reach and only use ϕ. Figure 2 and Figure 3 plot an MDS projection of the representations in 2D for L2 and Multilinear representations respectively. Our choice of MDS projection is motivated by its distance preserving nature even after projection.

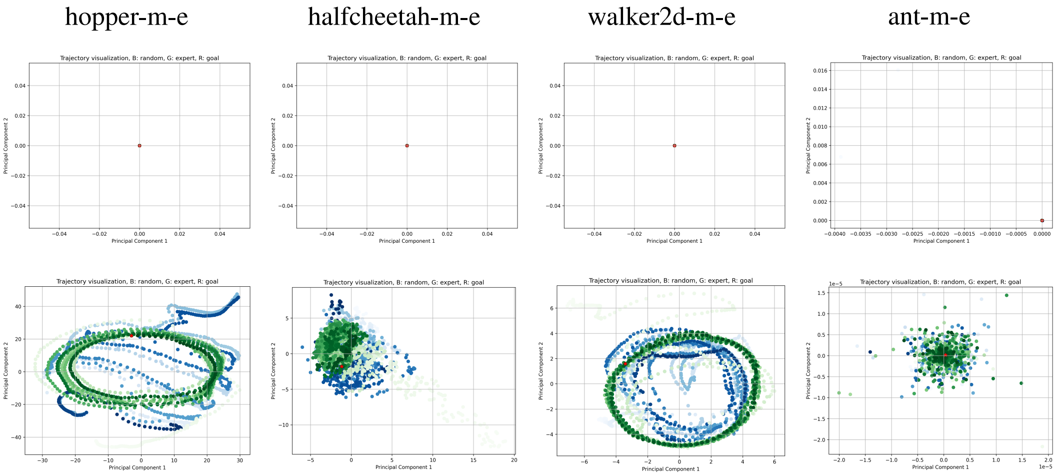

Figure 2: MDS plot of in-distribution representations with L2 bottleneck: We sample an expert trajectory that is not a part of the medium-expert data

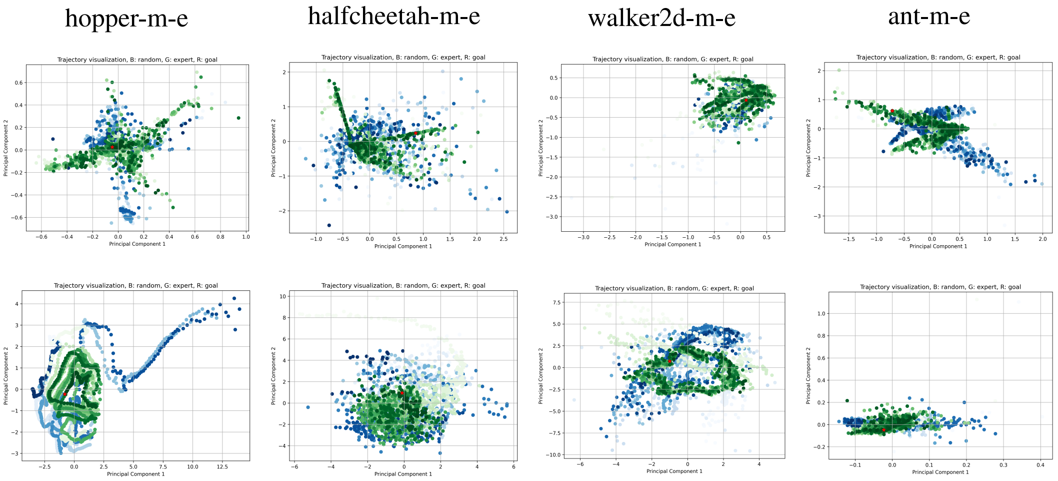

Figure 3: MDS plot of in-distribution representations with Multilinear bottleneck: We sample an expert trajectory that is not a part of the medium-expert data

In Figures 2, we observe VIP to collapse representations for most L2 tasks likely due to failing to implement a necessary embedding collision objective in its practical algorithm. Our representations for Hopper and Walker2d environments are most interpretable - The representation of expert trajectory forms a complete loop indicating the periodic pattern of states these environments encounter in an expert trajectory.

6.2 Optimal Value Functions with Representation Bottleneck

We evaluate the ability of VIP and to learn optimal value functions which is a direct indicator of the quality of representations learned. In this section, we sample an in-distribution expert trajectory - one expert trajectory that is in-distribution but does not exist in the medium-expert datasets and a random trajectory. We then fix the end state of the expert trajectory as the goal and plot the learned value function at every state of the trajectory.

In Figure 4 and Figure 5 we observe that 's value reflect the correct patterns we expect in the expert trajectory. For Hopper and Walker2d environments where the agent follows a periodic trajectory and observes similar state multiple times, the learned value function captures this behavior. For HalfCheetah and Ant, which are not periodic in nature as a result of choice of their observation space, the value function increases until the goal is reached. In contrast, VIP's value function is noisy and does not reflect the expected trends.

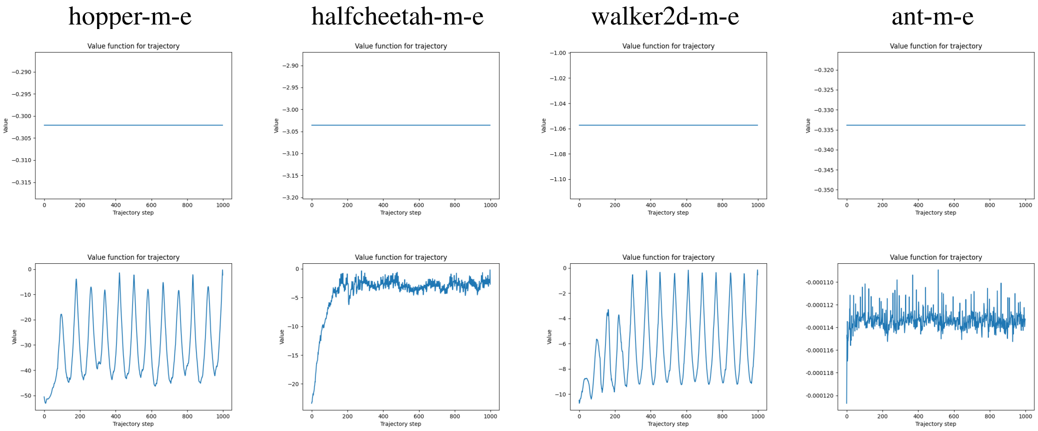

Figure 4: Optimal value function prediction on in-distribution trajectories under a bottlenecked L2 representation. VIP representations collapses causing the value function to collapse. learns meaningful value functions.

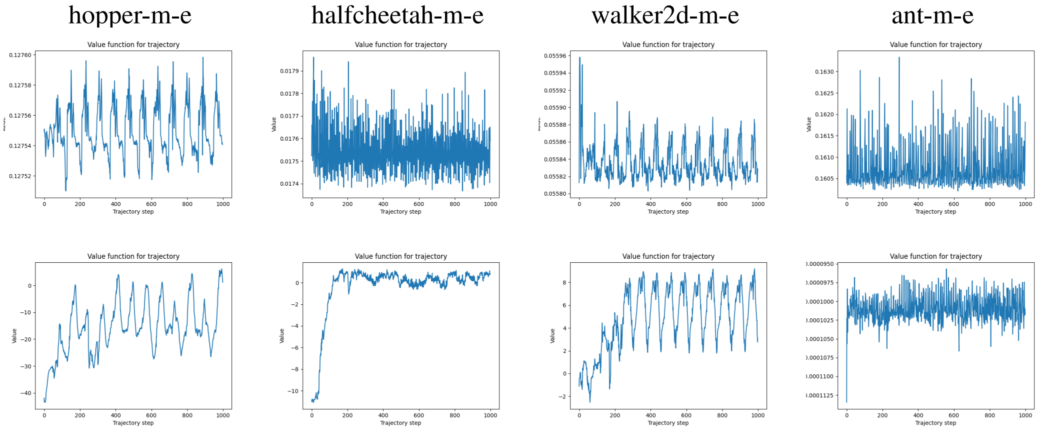

Figure 5: Optimal value function prediction on in-distribution trajectories under a bottlenecked Multilinear representation. VIP representations do not collapse but learn incorrect value functions for HalfCheetah and Ant environments. learns meaningful value functions.

7 Conclusion

Learning a general-purpose representation of the world for reinforcement learning has the potential to pave way to a foundational model for robotics. This work discusses the limitations of prior work, Value Implicit Pretraining (VIP), that learns such representations by training optimal goal-reaching value functions. Our core insight is assumptions made in VIP does not allow it to learn optimal value functions and hence the right representations. Our work proposes a clean, simple and effective alternative that overcomes these limitations and proposes an action-free objective for GCRL. To this end, we support our claims by studying the learned representations on simulated tasks from the D4RL benchmark.

References

Rishabh Agarwal, Marlos C Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning.

Siddhant Agarwal, Aaron Courville, and Rishabh Agarwal. Behavior predictive representations for generalization in reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021. URL `https://openreview.net/forum?id=b5PJaxS6Jxg`.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pp. 11321–11339. PMLR, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pp. 1094–8. Citeseer, 1993.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pp. 23301–23320. PMLR, 2023.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2024.

Liyuan Mao, Haoran Xu, Weinan Zhang, and Xianyuan Zhan. Odice: Revealing the mystery of distribution correction estimation via orthogonal-gradient update. *arXiv preprint arXiv:2402.00348*, 2024.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations, 2021.

Rutav M. Shah and Vikash Kumar. RRL: resnet as representation for reinforcement learning. *CoRR*, abs/2107.03380, 2021. URL `https://arxiv.org/abs/2107.03380`.

Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and Scott Niekum. Score models for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2311.02013*, 2023a.

Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual rl: Unification and new methods for reinforcement and imitation learning. *arXiv preprint arXiv:2302.08560*, 2023b.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, et al. Bridgedata v2: A dataset for robot learning at scale. *arXiv preprint arXiv:2308.12952*, 2023.

Yi Zhao, Wenshuai Zhao, Rinu Boney, Juho Kannala, and Joni Pajarinen. Simplified temporal consistency reinforcement learning, 2023.

Algorithm 2 DERAIL

```
Init $V_\phi(s, g)$, conservatism $\lambda$
Let $\mathcal{D} = \hat{\rho} = \{(s, a, s')\}$ be an offline dataset.
for $t = 1..T$ iterations do
    Train $V_\phi$ via Orthogonal gradient update on Eq. 13
end for
return $\phi$
```