Denoised Predictive Imagination: An Information-theoretic approach for learning World Models

Vedant Dave[*]
Cyber-Physical-Systems Lab
Montanuniversität Leoben, Austria

Elmar Rueckert
Cyber-Physical-Systems Lab
Montanuniversität Leoben, Austria

Abstract

Humans excel at isolating relevant information from noisy data to predict the behavior of dynamic systems, effectively disregarding non-informative, temporally-correlated noise. In contrast, existing reinforcement learning algorithms face challenges in generating noise-free predictions within high-dimensional, noise-saturated environments, especially when trained on world models featuring realistic background noise extracted from natural video streams. We propose a novel information-theoretic approach that learn world models based on minimising the past information and retaining maximal information about the future, aiming at simultaneously learning control policies and at producing denoised predictions. Utilizing Soft Actor-Critic agents augmented with an information-theoretic auxiliary loss, we validate our method's effectiveness on complex variants of the standard DeepMind Control Suite tasks, where natural videos filled with intricate and task-irrelevant information serve as a background. Experimental results demonstrate that our model outperforms nine state-of-the-art approaches in various settings where natural videos serve as dynamic background noise. Our analysis also reveals that all these methods encounter challenges in more complex environments.

1 Introduction

A major open problem in Reinforcement learning (RL) is to learn the dynamics and control policies from the high-dimensional observations such as images [Ha and Schmidhuber, 2018, Lillicrap et al., 2016, Hafner et al., 2020a, 2021a, Hansen et al., 2022]. Conventionally, it is assumed that the observations in the environment, often derived through hand-engineered features, consist exclusively of task-relevant information. This allows RL algorithms to operate in a controlled setting with optimal efficiency, primarily due to the absence of exogenous noise (unrelated or uncontrollable external variables such as weather variations or random background movements), that could potentially hinder the learning process.

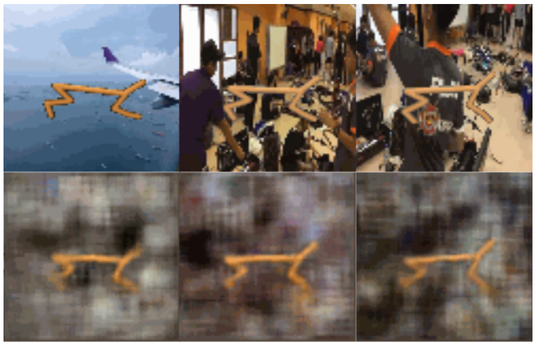


Figure 1: **Top Row:** Ground truth data from a random sequence. **Bottom Row:** Reconstruction from DPI.

[*]Corresponding Author: vedant.dave@unileoben.ac.at

17th European Workshop on Reinforcement Learning (EWRL 2024).

In the real world, the landscape is vastly different, brimming with a plethora of information, much of which is irrelevant to a specific task. The challenge lies in accurately identifying task-relevant information and avoid the modeling of temporally correlated dynamics of the background noise. Prior RL methodologies [Yarats et al., 2021, Hafner et al., 2020a, Ha and Schmidhuber, 2018] that derive representations directly from observations, often integrate task-irrelevant information into their representations. They struggle to disentangle the noise from relevant information, unnecessarily modeling noise dynamics, leading to sub-optimal performance under noise (see Table 1).

The process of computing representations relies on the past inputs, while the imagination and exploration are directed towards future [Hafner et al., 2020b]. Our objective is to develop a cohesive perspective on how an agent formulates its current representation after observing past input and before observing future. Could it be feasible to model this process as an information flow, transitioning from past to future, mediated by the current state?

We introduce Denoised Predictive Imagination (DPI), a model-based reinforcement learning approach that leverages information theory to learn robust and meaningful representations. DPI models Predictive Information [Bialek and Tishby, 1999], the mutual information between the past and the future, and employs the Information Bottleneck principle [Tishby et al., 2000] to derive a compact representation of the current state from historical observations, while preserving maximal predictive information about future outcomes. Essentially, DPI focuses on learning a concise abstraction of the system dynamics and leverages it to learn control policies and generate noise-free future predictions. This is achieved through deriving an objective integrating two central ideas: minimization of mutual information about past and the maximization of predictive ability for future. This dual objective consists of two contrastive losses and is formulated as a Lagrangian optimization problem. While in this paper we focus on the algorithmic derivation and the performance of DPI, the information theoretic nature of it enables future investigations of generalization, stability and robustness aspects. The primary contributions of our work are as follows:

1. This work is the first to demonstrate that denoised state representations can be effectively derived through the preservation of predictive information.
2. By implicitly integrating various methodologies from prior studies, we present a theoretical generalized framework for world model learning within the context of bottleneck methods.
3. DPI surpasses nine existing approaches across the majority of modified DeepMind control (DMC) tasks, additionally showing superior noise-free prediction capabilities alongside dynamic learning.

2 Related Work

In this section, we delve into related work on reinforcement learning from visual input, focusing specifically on model-based approaches and representation learning concepts. For a more comprehensive discussion, refer to the Supplementary Material.

Learning Control from pixels with distractors. Recent developments in model-based RL [Zhang et al., 2021, Ma et al., 2021, Nguyen et al., 2021, Fu et al., 2021, Bai et al., 2021, You et al., 2022, Bharadhwaj et al., 2022, Wang et al., 2022, Islam et al., 2022, Tomar et al., 2023, Liu et al., 2024] have put forward a variety of innovative ideas aimed at extracting relevant information from observations. Contrastive Variational Reinforcement Learning (CVRL, Ma et al. [2021]) aims at maximises the MI between observations and representations i.e. $I(o_t, z_t)$, which is exactly similar to our objective "Predictive Observation Model" in Equation 3 (except we consider it for all the timesteps and not just a single instance) and leverages InfoNCE contrastive loss [Oord et al., 2018] to optimise the objective. However, it does not address any objective related to generating noise-free dynamics or predictions, which can be observed via the reconstruction results in the original paper [Ma et al., 2021] (Figure 3, Page 7). MIRO [Ding et al., 2020] is another method that bears a close resemblance to CVRL. However, unlike CVRL, MIRO focuses on maximizing the mutual information (MI) between the state and observation, conditioned on the given action and constrained by dynamic predictions. Deep Bisimulation for Control (DBC, Zhang et al. [2021]) learns control policies by learning representations of the states that preserve the bisimulation metric. Temporal Predictive Coding (TPC, Nguyen et al. [2021]) shares conceptual similarities with our approach, striving to eliminate temporal noise while focusing only on the relevant aspects. The goal of TPC is to maximize the MI between future latent codes and the combination of prior latent codes and action tuples. This objective is achieved through contrastive learning, which exhibits a mathematical resemblance to Equation 10. More recent methods such as Task Informed Abstractions (TIA, Fu et al. [2021]) maintain two separate latent models, one for tasks and another for distractors, bifurcating noise and signal. TIA falters in achieving better rewards when the grayscale background is replaced with RGB (see the experimental section). Iso-Dreamer [Pan et al., 2022] learns inverse dynamics model to understand the controllable and non-controllable state-action relationship. It then aims to decouple these dynamics by rolling out their latent representations into the future to understand how these dynamics influence current behavior. Our work bypasses the need for explicitly defining these types of model rules and instead builds on a general information-theoretic model wherein these types of features implicitly emerge.

3 Notation and Preliminaries

Reinforcement Learning. An agent operates in a Markov Decision Process (MDP), which is characterised by a tuple $\mathcal{M} = (\mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, consisting of the observation space $\mathcal{O}$ with observations o (we interchangeably use "states" and "observations"), action space $\mathcal{A}$ with actions a, transition dynamics $\mathcal{P}$, Reward space $\mathcal{R}$ and discount factor $\gamma \in [0, 1)$. The encoder $\phi(z|o)$ produces a latent representation z from observations, and then the policy $\pi(a|z)$ decodes this latent representation into actions. The goal of RL is to learn a policy $\pi^*(a|z)$ that maximizes the expected cumulative discounted rewards $\mathcal{J}_\pi = argmax_\pi \mathbb{E}_p \big[\sum_t \gamma^{t-1} r_t \big]$.

Predictive Information. Predictive Information (PI) is a quantity that measures how much our observations from the past can inform us about the future [Bialek and Tishby, 1999] . Mathematically, it can be defined as the mutual information (MI) between the past (x_{past}) and the future (x_{future}), denoted as $I(x_{past};\, x_{future})$. Assuming temporal invariance (any fixed time length is expected to have the same entropy), PI becomes a subextensive quantity, as expressed by $\lim_{T\to\infty} I(T)/T = 0$, where $I(T)$ is the predictive information over a time window of length 2T (with T steps of the past predicting T steps into the future), see Equation 3.1 in [Bialek et al., 2001]. As the time frame increases, the past contains a diminishing predictive value for the future. In order to capture only the necessary information from x_{past} for predicting x_{future}, a compressed representation of x_{past} is required.

Information Bottleneck. For learning this compressed representation, we utilize the Information Bottleneck (IB) principle [Tishby et al., 2000]. IB aims at learning a representation z that aims to optimally compress the information provided by the input $x \in X$, i.e. minimize $I(x; z)$, while still maintaining enough knowledge to predict the outcome $y \in Y$, i.e. maximize $I(z; y)$. This objective is unified with the inclusion of a Lagrangian multiplier and formalized as $max\ I(z; y) - \beta I(x; z)$. The parameter β controls the information flow from the input x to the latent representation, balancing the trade-off between information preservation and compression.

4 Denoised Predictive Imagination

Denoised Predictive Imagination (DPI) is an information theory-based approach, that encapsulates the notions of predictive information and the information bottleneck. This core idea enables the learning of a compressed representation from high-dimensional observations, distilling task-relevant details from past observations, and leveraging this refined knowledge for future predictions while effectively filtering out noise. We hypothesise that the current state should encapsulate the requisite and meaningful information essential to perform the task. If the information is insufficient, the latent representations may fail to capture all the task-relevant information, leading to sub-optimal learning outcomes. On the other hand, if we incorporate an overabundance of information, our representations could become encumbered with noise-related artifacts that results in a dilution of task-relevant data and in a performance decrease.

We denote the latent representations for the past observations by o_{t-}, current observation by o_t, and the future observations by o_{t+}. We use z_{t-}, z_t and z_{t+} respectively for the latent space. For consistency and clarity, we establish that the episode initiates at time $t = 1$ and terminates at the horizon $t = T$. The objective is to encode observations $o_{\leq t} = (o_{t-}, o_t)$ into latent representations $z_{\leq t} = (z_{t-}, z_t)$, transform them to next state representations z_{t+}, and decode into future observations o_{t+} (Figure 2). Consequently, this process creates a two-fold bottleneck: one while transforming observations into latent representations and vice-versa ($o_t \leftrightarrow z_t$), and another when acquiring the latent representation itself from other latent representations ($z_{t-1} \rightarrow z_t \rightarrow z_{t+1}$). In this context, our transition function can be conceptualized as a model operating simultaneously as an encoder and a decoder, encoding z_t from z_{t^-} and decoding z_t to yield z_{t^+}, with bottleneck being z_t.

Intuitively, we obtain task-relevant information from raw observations into our latent representations by minimising mutual information $I(o_{\leq t}; z_{\leq t})$ while maximising the mutual information $I(o_{\geq t}; z_{\geq t})$, which preserves the predictive information for the reverse scenario. When expressed in Lagrangian formulation, we obtain,

$$\min \; I(o_{\leq t}; z_{\leq t}) - \beta_1 I(o_{\geq t}; z_{\geq t})). \tag{1}$$

In order to learn temporal abstractions and compressed representations from a sequence of past states and acquire relevant predictions, we employ the principle of Information Bottleneck. We apply a Lagrangian on the latent space with the aim of minimising $I(z_{\leq t})$ and maximising $I(z_{\geq t})$,

$$\min \; I(z_{\leq t}) - \beta_2 I(z_{\geq t}). \tag{2}$$

Merging objectives from equation (1) and (2), we obtain a unified Lagrangian optimizing problem,

$$\min \; \Big[\underbrace{I(o_{\leq t}; z_{\leq t})}_{\substack{\text{Historical} \\ \text{observation model}}} + \underbrace{I(z_{\leq t})}_{\substack{\text{Historical latent} \\ \text{space dynamics}}} \Big] - \Big[\underbrace{\beta_1 I(o_{\geq t}; z_{\geq t})}_{\substack{\text{Predictive} \\ \text{observation model}}} + \underbrace{\beta_2 I(z_{\geq t})}_{\substack{\text{Predictive latent} \\ \text{space dynamics}}} \Big],$$

where β_1 and β_2 are the Lagrangian multipliers. This implies that the problem can be optimised by minimizing the upper bound associated with the past, as represented by the first two terms, and simultaneously maximizing the lower bound related to the future, embodied in the final two terms. The objective of our DPI considers action dependencies implicitly through the latent space representations, $p(z_t | z_{t^-}, a_{t^-})$, thereby reflecting the innate characteristics of system transitions. This compatibility with RL principles facilitates a seamless integration of our approach into existing RL algorithms, where DPI can serve as an auxiliary function, significantly enhancing the learning of representations. Due to space limitations, all subsequent derivations and details are in the Supplementary Material (Section 1).

4.1 State Space Model

We use the state-space model described in Figure 2 with,

$$\begin{array}{lr} \text{Encoder Representation:} & z_t \sim p_\varphi(z_t \mid o_t) \\ \text{Transition dynamics:} & z_{t+1} \sim q_\theta(z_{t+1} \mid z_t, a_t, h_t) \\ \text{Observation model:} & o_t \sim r_\psi(o_t \mid z_t) \\ \text{History model:} & h_t \sim p(h_t \mid h_{t-1}, a_{t-1}). \end{array} \tag{3}$$

The conditional $p(h_t \mid h_{t-1}, a_{t-1})$ denotes the history model, that encapsulates the past variables into a single history variable i.e.,

$$h_t = \{z_{t-1}, a_{t-1}, ..., z_1, a_1\} = \{z_{t-1}, a_{t-1}, h_{t-1}\}. \tag{4}$$

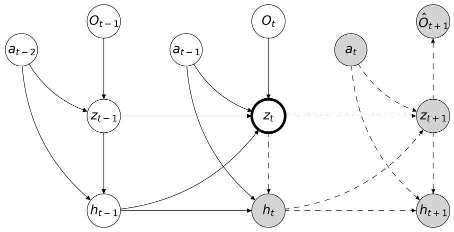


Figure 2: **State-space model.** The variable z_t acts as a bottleneck for the model, serving as a critical link between the historical (white circles) and predictive elements (grey circles). Solid edges designate the inputs required for inference, while the dotted edges represent the generative components.

4.2 Minimising the upper bound of the Past Mutual Information

This subsection discusses the minimization of the first two terms in the Lagrangian of DPI in Equation 3.

Upper bound of historical latent space dynamics. We aim at minimising the tractable upper bound on the mutual information $I(z_{\leq t})$. The mutual information can be represented as,

$$I(z_1; ...; z_t) = \mathbb{E}_{p(z_1,...,z_t)} \left[\log \frac{p(z_1, ..., z_t)}{\prod_{k=1}^{t} p(z_k)} \right],$$

We incorporate actions into the model by introducing a conditional probability distribution $p(z_{t^-}, z_t|a_{t^-})$,

$$I(z_{1:t}) = \mathbb{E}_{p(z_{1:t},a_{1:t-1})}\left[\log \frac{p(z_{1:t})p(z_{1:t}|a_{1:t-1})}{p(z_{1:t}|a_{1:t-1})\prod_{k=1}^{t} p(z_k)}\right] \leq \mathbb{E}_{p(z_{1:t},a_{1:t-1})}\left[\log \frac{p(z_{1:t}|a_{1:t-1})}{\prod_{k=1}^{t} p(z_k)}\right]. \tag{5}$$

Utilising the chain rule in conditional probability and for every t, substituting $\{z_{t-1}, a_{t-1}, h_{t-1}\}$ as h_t like Equation (4), we can write Equation (5) as

$$I(z_{1:t}) \leq \sum_{k=1}^{t-1} \mathbb{E}_{p(z_k,a_k)}\left[\log \frac{p(z_{k+1}|z_k, a_k, h_k)}{p(z_{k+1})}\right] = \sum_{k=1}^{t-1} I\big(z_{k+1}; z_k, a_k, h_k\big). \tag{6}$$

In essence, this implies that we can optimize the mutual information between the past latent representations and the present state's representation by minimising the upper bound of the MI for each individual, independent transition in a Markovian manner.

For the purpose of minimizing this upper bound, we employ Contrastive Log-ratio Upper Bound of Mutual Information (CLUB, Cheng et al. [2020]), where the core idea is to estimate the MI through the difference of conditional probabilities for positive and negative sample pairs. Since the conditional distribution $p(z_{k+1}|z_k, a_k, h_k)$ is intractable, the upper bound of $I(z_{k+1}; z_k, a_k, h_k)$ cannot be directly minimized. As a consequence, we introduce a variational distribution $q(z_{k+1}|z_k, a_k, h_k)$, serving essentially as the transition function of the model, parameterised by θ, to approximate the upper bound of mutual information,

$$I(z_{k+1}|z_k, a_k, h_k) = \frac{1}{N}\sum_{i=1}^{N}\left[\log \hat{q}_\theta - \frac{1}{N}\sum_{j=1}^{N}\log \hat{q}_\theta\right] = I_{\text{CLUB}}\ , \tag{7}$$

where $\hat{q}$ denotes $q_\theta(z^i_{k+1}|z^i_k, a^i_k, h^i_k)$, i.e. the i-th sample at k-th timestep. We obtain the negative sample pair $(z'_{k+1}, (z_k, a_k, h_k))$ via random shuffling.

Upper bound of the historical observation model. As in the previous section, it can be shown that an upper bound for $I(o_{1:t}, z_{1:t})$ can be derived by introducing the conditional distribution $p(z_{t^-}, z_t|a_{t^-})$,

$$I(o_{1:t}; z_{1:t}) \leq \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log \frac{p(z_{1:t}|o_{1:t})}{p(z_{1:t}|a_{1:t-1})}\right].$$

Taking this further, we employ the same tractable variational distribution drawn from our transition function,

$$I(o_{1:t}; z_{1:t}) \leq \sum_{k=1}^{t-1} \mathbb{E}_{p(z_k,o_k,a_k)}\left[\log \frac{p(z_{k+1}|o_{k+1})}{q_\theta(z_{k+1}|z_k, a_k, h_k)}\right] = I_{\text{LTC}}\ . \tag{8}$$

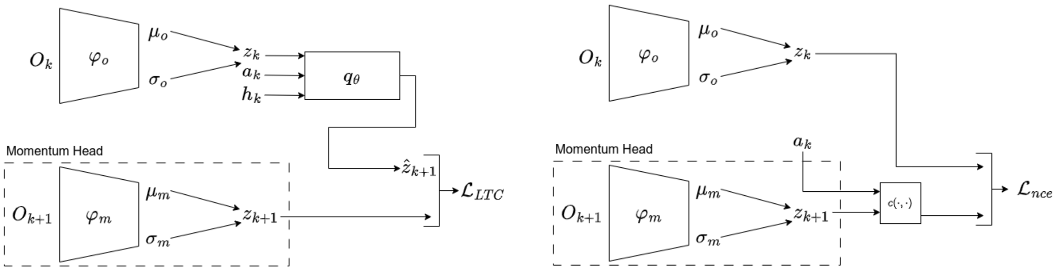
This term is an upper-bound for $I(o_{1:t}, z_{1:t})$, quantifying the ratio between the latent representation derived from the encoder and the transitioning state obtained from a past representation when a specific action is applied. Intuitively, this constrains the latent dynamical model (transition function) to diverge minimally from the latent representations obtained from the observation encoder. Hence, we refer to this term as the Latent Consistency Loss $\mathcal{L}_{\text{LTC}}$.

4.3 Maximising the lower bound of the Predictive Mutual Information

This subsection discusses the maximization of the last two terms in the Lagrangian of DPI in Equation 3.

Lower bound of the predictive latent space dynamics. In order to obtain the lower bound on this MI term, we factorise the transition model by applying the chain rule,

$$\begin{aligned} I(z_{t:T}) &= \mathbb{E}_{p(z_{t:T})}\left[\log \frac{p(z_{t:T})}{\prod_{k=t}^{T} p(z_k)}\right] = \mathbb{E}_{p(z_{t:T})}\left[\log \prod_{k=t}^{T-1} \frac{p(z_k|z_{k+1:T})}{p(z_k)}\right] \\ &\geq \sum_{k=t}^{T-1} \mathbb{E}_{p(z_k,a_k)}\left[\log \frac{p(z_k|z_{k+1}, a_k)}{p(z_k)}\right] = \sum_{k=t}^{T-1} I\big(z_{k+1}, a_k; z_k\big). \end{aligned} \tag{9}$$



(a) Latent transition consistency loss Model

(b) NCE loss Model

Figure 3: Representation of models used for calculating auxiliary losses (a) LTC loss $\mathcal{L}_{LTC}$ and NCE loss ($\mathcal{L}_{NCE}$). Encoder and target encoder parameters are defined as φ_0 and φ_m respectively. a) Once the current representation is obtained, it is passed through the transition function q_θ to obtain the next latent representation, from which the $\mathcal{L}_{LTC}$ is finally calculated (Algorithm 2 in Supplementary material). b) Next latent representation and current action is passed via concatenation function c to obtain unified representation, then compared with current state representation via contrastive learning.

The mutual information objective $I(z_{k+1}, a_k; z_k)$ can be decomposed using the chain rule for mutual information, yielding $I(z_k; z_{k+1}) + I(z_k; a_k|z_{k+1})$. The first component, solely depends on state-transitions. It is closely related to the predictive coding objective [Oord et al., 2018, Anand et al., 2019]. Omitting actions could impair the model's capability to determine the optimal actions [Rakelly et al., 2021]. The second term can be represented in terms of conditional entropy as $H(a_k|z_k) - H(a_k|z_k, z_{k+1})$. The term $H(a_k|z_k, z_{k+1})$ effectively characterizes the entropy of the inverse dynamics, conceptually aligns closely with an extensive spectrum of prior studies that have focused on exploration and unsupervised learning of representations [Zhang et al., 2018, Pathak et al., 2017, Chandak et al., 2019, Bharadhwaj et al., 2022]. Also, this term is the empowerment objective used in InfoPOWER [Bharadhwaj et al., 2022]. From an intuitive perspective, inverse models operate as an agreement mechanism between the actual and the ground truth action representations. This mechanism enables the representation to capture only those aspects of the state that are essential for predicting the action, thereby discarding potentially irrelevant information. The MI term in Equation 9 can be viewed as a combined objective that optimises state transitions with the regularization of action representations.

For optimising this lower bound, we will utilise contrastive learning [Oord et al., 2018], which yields a variational lower bound of the mutual information in Equation 9. Strategies employed by [He et al., 2020, Laskin et al., 2020] relies on data augmentation to generate positive and negative samples. Contrary to them, we take inspiration from Bai et al. [2021] that incorporate policy transitions to obtain these samples. Positive samples are directly acquired by sampling transitions (z_t, a_t, z_{t+1}), while the construction of negative samples involves randomly sampling z_t^* and concatenating it with (a_t, z_{t+1}). As a result, we produce samples (z_t^*, a_t, z_{t+1}) that deviate from the transition dynamics. Thus we obtain MI objective as,

$$I(z_{k+1}, a_k; z_k) \geq \mathbb{E}_{p,N}\left[\log \frac{e^{\sigma(z_k,a_k,z_{k+1})}}{\sum_{z_k^* \in N^- \cup z_k} e^{\sigma(z_k^*,a_k,z_{k+1})}}\right] \triangleq I_{\text{NCE}} \ , \tag{10}$$

where N is the set of negative samples and σ is the score function. Score function distinguishes between positive samples (those following the actual transition dynamics) and negative samples (those deviating from these dynamics). It providing high score to the positive examples and low score to the negative examples. It tells how well an action a_k leads to a transition from a latent state z_k to a subsequent latent state z_{k+1}. This evaluation is based on the degree to which the action and the resultant state change are congruent with the expected dynamics of the system. We opt for bilinear products as our score function [Oord et al., 2018, Laskin et al., 2020, Henaff, 2020], which is mathematically defined as $\sigma(z_k, a_k, z_{k+1}) = c(a_t, z_{t+1})^T \mathcal{W} z_t$. The concatenation function $c(\,\cdot\,,\,\cdot\,)$ is parameterised by a neural network that merges the action a_t with the subsequent latent state z_{t+1} into a single vector (as shown in Figure 3b) and $\mathcal{W}$ is the learnable weight matrix.

Lower bound of the predictive observation model. Directly maximizing $I(z_{t,t^+}; o_{t,t^+})$ is infeasible due to its marginal's intractability. Similar to Alemi et al. [2017], we propose to optimise a lower bound on our MI,

$$I(z_{t:T}; o_{t:T}) = \mathbb{E}_{p(z_{t:T}, o_{t:T})} \left[\log \prod_{k=t}^{T} \frac{p(o_k|z_k)}{p(o_k)} \right] \geq \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)} \left[\log \frac{r_\psi(o_k|z_k)}{p(o_k)} \right],$$

where where $p(o_k|z_k)$ is an intractable conditional distribution and $r_\psi(o_k|z_k)$ is a tractable variational decoder, represented by a neural network with parameters ψ. We rule out the entropy term as it is independent of our optimization procedure,

$$I(z_{t:T}; o_{t:T}) = \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)} \Big[\log \; r_\psi(o_k|z_k) \Big] = I_{\text{Rec}} \; . \tag{11}$$

I_{Rec} can be interpreted as the log-likelihood of the observations given the state encodings.

4.4 Combined Objective

Our optimization strategy can be unified into a single objective function as,

$$\min_{\theta, \psi, \phi, \mathcal{W}} \mathcal{L}_{DPI} = \; [\alpha_1 I_{LTC} + \alpha_2 I_{CLUB}] - [\beta_1 I_{Rec} + \beta_2 I_{NCE}]. \tag{12}$$

The two losses, I_{LTC} and I_{Rec}, are responsible for the representations from the encoder and decoder respectively, while the other two terms, I_{CLUB} and I_{NCE}, formulated as a contrastive loss, control the representations of the transition functions. They are jointly optimized.

4.5 Practical Implementation with Soft-Actor Critic

We jointly train DPI and SAC, an off-policy model-free reinforcement learning method, by incorporating Equation (12) as an auxiliary objective while training the algorithm (Supplementary Material Section 3.1). The transition model, accounting for latent dynamics, is designed to capture the inherent stochasticity of the transitions. It is parameterised with a neural network that returns a Gaussian distribution defined by its mean and variance. The Observation model implemented as a Deconvolutional Neural Network [Zeiler et al., 2010]. The History model is implemented as a Gated Recurrent Unit (GRU, Cho et al. [2014]). We utilize a stochastic encoder to obtain representations from the images [Eysenbach et al., 2021, Theis and Agustsson, 2021], parameterised by φ. For encoding subsequent observations, we leverage an exponential moving average of the online network parameters, denoted as φ_m [He et al., 2020]. We utilise the same principle for latent targets [Hansen et al., 2022] for transition function, as it should ensure more stable learning process, accommodating any potential fluctuations in the learning (Figure 3a). The complete algorithm with SAC is described in the Supplementary material.

5 Experiments

In this section, we conduct a thorough empirical assessment of the proposed DPI method on the DeepMind control suite (DMC, Tassa et al. [2018]) in various settings and compare it with existing state-of-the-art approaches. We evaluate three distinct types of environments: (i) Standard environment with a static background, (ii) Natural environment with video-based, real-world backgrounds, and (iii) Random environment with varying backgrounds in each frame. These settings can be viewed as introducing increasing levels of noise in the visual observations, with the stochasticity of the noise escalating in each scenario. It is important to highlight that the noise considered in our setup is specifically related to the background of the observation, while the controlled part of the image remains unaffected. This type of noise is particularly relevant in practical applications, where background clutter or irrelevant visual information can significantly impact the performance of RL agents. The ability to maintain robust policy learning in the presence of such noise is crucial for deployment in real-world environments, where the visual scene is often complex and dynamic. To underline the significance of each element in the model, we conclude this section with an ablation study.

5.1 Environment Settings

For all three environments, we conducted experiments on six DMC tasks: Cheetah Run, Walker Walk, Cartpole Swingup, Reacher Easy, Pendulum Swingup and Cup Catch. These robot control tasks pose different challenges, such as sparse rewards, contacts and complex dynamics. For the standard settings, no perterbutations are applied to the observations. The observations are RGB images of the size 84 × 84 × 3. By incorporating the ground plane, a substantial portion of the background image is obscured, thereby simplifying the task at hand. Thus, the ground plane is eliminated to maximize the utilization of the background image. These natural videos are incorporated from Kinetics 400 dataset [Kay et al., 2017] at random. We used videos from random categories compared to the simplified challenge in DBC [Zhang et al., 2021] who only considered the driving category. Contrary to the predominant use of grayscale images in benchmarking, we employing RGB videos in the background. We independently sampled 100 videos separately for training and testing. More information about the background noise is provided in the Supplementary Material (Section 5.1).

Table 1: Rewards in Natural Environment Background Settings

Method	**CR**	**WW**	**CS**	**PS**	**RE**	**CC**
DBC	122 ± 4	74 ± 41	181 ± 48	**26 ± 46**	**305 ± 470**	0 ± 0
De-MDPs	71 ± 18	113 ± 26	73 ± 2	0 ± 0	83 ± 33	0 ± 0
Dreamer	42 ± 9	68 ± 31	109 ± 46	0 ± 0	129 ± 188	68 ± 31
Dreamer-V2	118 ± 51	39 ± 29	137 ± 78	0 ± 0	0 ± 0	0 ± 0
SPR	45 ± 59	37 ± 6	150 ± 21	7 ± 10	100 ± 78	99 ± 1
TIA	20 ± 14	80 ± 52	118 ± 11	0 ± 0	115 ± 161	237 ± 411
TPC	42 ± 37	30 ± 9	106 ± 27	25 ± 35	16 ± 3	237 ± 334
VSG	56 ± 14	232 ± 43	139 ± 10	0 ± 0	12 ± 17	0 ± 0
Iso-Dreamer	10 ± 4	250 ± 48	99 ± 50	0 ± 0	12 ± 3	0 ± 0
DPI (Ours)	**263 ± 11**	**454 ± 60**	**658 ± 62**	**40 ± 57**	**308 ± 222**	**332 ± 576**

The table illustrates the rewards obtained in natural background settings across a variety of tasks. The best or comparable method is present in bold. Shorthands: CR - Cheetah Run, WW - Walker Walk, CS - Cartpole Swingup, PS - Pendulum Swingup, RE - Reacher Easy, CC - Cartpole Swingup, De-MDPs - Denoised MDPs.

5.2 Baselines and Implementation details

In this evaluation, we compare our approach to a selection of nine most-closely related approaches i.e. Dreamer [Hafner et al., 2020a], Dreamer-V2 [Hafner et al., 2021b], Task-informed Abstractions (TIA, Fu et al. [2021]), Denoised MDPs [Wang et al., 2022], Deep Bisimulation for Control (DBC, Zhang et al. [2021]), Self-Predicting Representations (SPR, Schwarzer et al. [2021]), Variational Sparse Gating (VSG, Jain et al. [2022]), Iso-Dreamer Pan et al. [2022] and Temporal Predictive Coding (TPC, Nguyen et al. [2021]). These selected methods are distinguished by their superior performance and accompanied by publicly accessible source code. The task return is examined every 1000 steps. For all baseline methods, we employed the optimal set of hyperparameters as indicated in the respective papers. Each task is executed with three different seeds for each model. Detailed explanations of these methods and of the implementations can be found in the Supplementary Material (Section 4).

5.3 Results in Standard settings

The performance of all the evaluated methods in the standard DMC environment is illustrated in the Supplementary Material (Section 5.4). DPI exhibits a degree of effectiveness in certain scenarios involving static backgrounds, although it does not consistently outperform all other methods.

5.4 Results in Natural Background settings

Figure 1 (Supplementary Material) illustrates the outcomes when employing natural backgrounds, wherein the background videos were not presented to the agent during its training phase. The main reasons for the degraded performance of most baseline methods was changing the background image to RGB. Dreamer struggles to accurately capture the agent's entire state, and inadvertently incorporates the irrelevant background noise into its representation (Supplementary Material Section 6). TIA, on the other hand, can only effectively distinguish the agent from the distractor when the background is rendered in grayscale. DBC's performance is on par with these methods, however, it does not achieve the performance that was reported by Zhang et al. [2021]. This discrepancy is largely due to the inclusion of RGB image in the background and authors' approach to use the same video for both training and testing, which hampers its capability to manage diverse distraction and restricts its generalization capability to unseen distractions. Similarly, TPC [Nguyen et al., 2021] and Denoised MDPs [Wang et al., 2022] underperformed due to its incapability to generalise to diverse unseen distractions. Our implementation utilises the authors' open-sourced code, with the sole adjustment being the introduction of additional videos. Contrary to these methods, DPI achieves better rewards in the top three environments in Table 1 (also see Figure 6 in Supplementary Material) demonstrates the superior performance of DPI across most environments. This is due to our bottleneck framework preserving predictive information during transitions, resulting in blurred backgrounds and enhanced agents. This highlights DPI's ability to encode task-relevant components, improving performance in complex and noisy environments (Reconstruction Results in Supp. Material Section 6). In Pendulum Swingup, Reacher Easy, and Cup Catch, performance is influenced by seed randomness. In Cup Catch, episodes starting with the cup in the holder lead to scores averaging 333 ± 576. In Reacher Easy, methods like VSG, DBC, Iso-Dreamer, and DPI rotate the arm instead of reaching, scoring higher inadvertently. This issue does not occur in Cheetah Run, Walker Walk, or Cartpole Swingup. In Cartpole Swingup, many methods rotate the cartpole, scoring 150-200, except DPI, which learns the intended swingup and balancing action.

Failure under Sparse rewards. Our approach excels in Dense reward scenarios (e.g., Cheetah run, Walker walk, Cartpole swingup). However, it struggles with sparse reward environments (Cup Catch and Pendulum Swingup) after 10^6 environment steps. The complexity of the task, when paired with the visual noise in the environment, presents a considerable challenge and surpasses the limits of current methodologies. In conclusion, the tasks that are inherently hard for model-based methods would remain hard for DPI. Significant improvements can be made for exploration in such environments.

5.5 Results in Random Background settings

In this experiment, every time instance features a unique background image, inducing maximum stochasticity in the environment. This experiment illustrates the preservation of temporally predictive information by DPI. As demonstrated in Figure 2 (Supplementary Material), for Cheetah run, DPI effectively isolates task-relevant features, managing to reconstruct only the agent against a randomized background. In Table 4, a comparative analysis is presented between DPI and nine baseline methodologies in Cheetah Run and Cartpole swingup environment. This shows superior performance of DPI over the baselines in natural background settings. The notable performance drop observed in Denoised MDPs [Wang et al., 2022] can be attributed to the introduction of varied and continually changing videos during the evaluation phase. It is likely that it has encountered the frames where the agent is not capable of segregating the relevant components from the non-relevant ones. This issue highlights a key limitation in its robustness and adaptability to varying environments.

6 Discussion and Conclusion

Our work demonstrates that our information-theoretic formulation suggests a pathway to segregate and represent task-relevant information in a noisy world, without explicitly modelling any rules of the MDPs. We also show that objectives related to maximising information on various variables, that are explicitly mentioned in other research [Bai et al., 2021, You et al., 2022, Lee et al., 2020b], implicitly emerge out from our theoretical formulation. In our analysis, all the methodologies exhibit strong performance in noise-free scenarios. When subjected to natural noise scenarios, characterized by real-world videos, DPI consistently either surpassed or equaled the best of nine baselines in performance. However, there's a noticeable path for improvement as every method encountered challenges in tasks dominated by sparse rewards (bottom row of Figure 1 Supplementary Material). Most notably, in random noise conditions, DPI does not face significant drop in performance and outperforms all other baseline methodologies. We assert that, while there have been notable contributions in the

segregation of controllable and non-controllable elements within scenes, the field is in dire need of algorithms that are capable of performing effectively in challenging and complex environments. This necessity is clearly underscored by our empirical analysis, which highlights the current limitations and underscores the importance of continued development in this area. While using vision alone may seem limited, integrating it with other types of data can lead to a more powerful multimodal approach. By combining vision with other modalities like proprioception, tactile, language etc., we can create a more robust multimodal approach that leverages the strengths of each modality [Becker et al., 2024, You and Liu, 2024, Peri et al., 2024, Dave et al., 2024, Yang et al., 2024, Wang et al., 2019, Yu et al., 2022]. The current method could be extended to incorporate these additional data sources, making it even more versatile.
Our method can be potentially combined with any existing RL model that performs exponentially well in noise-free environment. We believe that there is a great room for improving the performance of our model, e.g., by improving the model architecture for the encoding representations using Resnet like in Bai et al. [2021], by utilising experience replay sampling strategies like PER [Schaul et al., 2016], or by incorporating sophisticated exploration strategies for sparse environments [Laskin et al., 2020, You et al., 2022].

Acknowledgments and Disclosure of Funding

This project has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - No #430054590 (TRAIN).

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Chenjia Bai, Lingxiao Wang, Lei Han, Animesh Garg, Jianye HAO, Peng Liu, and Zhaoran Wang. Dynamic bottleneck for robust self-supervised exploration. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://openreview.net/forum?id=-t6TeG3A6Do.

Philipp Becker, Sebastian Mossburger, Fabian Otto, and Gerhard Neumann. Combining reconstruction and contrastive methods for multimodal representations in rl, 2024. URL https://arxiv.org/abs/2302.05342.

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=DfUjyyRW90.

William Bialek and Naftali Tishby. Predictive information. *arXiv preprint cond-mat/9902341*, 1999.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In *International Conference on Machine Learning*, pages 941–950. PMLR, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pages 1779–1788. PMLR, 2020.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL `https://aclanthology.org/W14-4012`.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Vedant Dave, Fotios Lygerakis, and Elmar Rueckert. Multimodal visual-tactile representation learning through self-supervised contrastive pre-training. *arXiv preprint arXiv:2401.12024*, 2024.

Yiming Ding, Ignasi Clavera, and Pieter Abbeel. Mutual information maximization for robust plannable representations. *arXiv preprint arXiv:2005.08114*, 2020.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Ben Eysenbach, Russ R Salakhutdinov, and Sergey Levine. Robust predictable control. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:27813–27825, 2021.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3480–3491. PMLR, 18–24 Jul 2021. URL `http://proceedings.mlr.press/v139/fu21b.html`.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pages 2170–2179. PMLR, 2019.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf`.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=S1lOTC4tDS`.

Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*, 2020b.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=0oabwyZbOu`.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=0oabwyZbOu`.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*. PMLR, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bklr3j0cKX`.

Riashat Islam, Manan Tomar, Alex Lamb, Hongyu Zang, Yonathan Efroni, Dipendra Misra, Aniket Rajiv Didolkar, Xin Li, Harm van Seijen, Remi Tachet des Combes, and John Langford. Agent-controller representations: Principled offline RL with rich exogenous information. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022. URL `https://openreview.net/forum?id=OpFzg-8y-o`.

Arnav Kumar Jain, Shivakanth Sujit, Shruti Joshi, Vincent Michalski, Danijar Hafner, and Samira Ebrahimi Kahou. Learning robust dynamics through variational sparse gating. In *Advances in Neural Information Processing Systems*, December 2022.

Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xCPJHtDB`.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Kuang-Huei Lee, Ian Fischer, Anthony Liu, Yijie Guo, Honglak Lee, John Canny, and Sergio Guadarrama. Predictive information accelerates learning in rl. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:11890–11901, 2020b.

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. URL `http://arxiv.org/abs/1509.02971`.

Yuren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. *Advances in Neural Information Processing Systems*, 36, 2024.

Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, and Igor Mordatch. Plan online, learn offline: Efficient learning and exploration via model-based control. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Byey7n05FQ.

Xiao Ma, Siwei Chen, David Hsu, and Wee Sun Lee. Contrastive variational reinforcement learning for complex observations. In *Conference on Robot Learning*, pages 959–972. PMLR, 2021.

Shahin Nasr, Ali Moeeny, and Hossein Esteky. Neural correlate of filtering of irrelevant information from visual working memory. *PLoS One*, 3(9):e3282, 2008.

Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In *International Conference on Machine Learning*, pages 8130–8139. PMLR, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *Advances in Neural Information Processing Systems*, 2022.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 2778–2787. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Skand Peri, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. *Conference on Robot Learning*, 2024.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which mutual-information representation learning objectives are sufficient for control?, 2021.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2016.

Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uCQfPZwRaUu.

Rui Shu, Tung Nguyen, Yinlam Chow, Tuan Pham, Khoat Than, Mohammad Ghavamzadeh, Stefano Ermon, and Hung Bui. Predictive coding for locally-linear control. In *International Conference on Machine Learning*, pages 8862–8871. PMLR, 2020.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

L. Theis and E. Agustsson. On the advantages of stochastic encoders. In *Neural Compression Workshop at International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2102.09270.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Manan Tomar, Riashat Islam, Sergey Levine, and Philip Bachman. Ignorance is bliss: Robust control via information gating. *arXiv preprint arXiv:2303.06121*, 2023.

Tongzhou Wang, Simon S. Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised mdps: Learning world models better than the world itself. In *International Conference on Machine Learning*. PMLR, 2022.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638, 2019.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26340–26353, June 2024.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10674–10681, 2021.

Bang You and Huaping Liu. Multimodal information bottleneck for deep reinforcement learning with multiple sensors. *Neural Networks*, 176:106347, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106347. URL `https://www.sciencedirect.com/science/article/pii/S0893608024002715`.

Bang You, Jingming Xie, Youping Chen, Jan Peters, and Oleg Arenz. Self-supervised sequential information bottleneck for robust exploration in deep reinforcement learning. *arXiv preprint arXiv:2209.05333*, 2022.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, JaeSung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, et al. Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*, 2022.

Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535, 2010. doi: 10.1109/CVPR.2010.5539957.

Amy Zhang, Harsh Satija, and Joelle Pineau. Decoupling dynamics and reward for transfer learning, 2018. URL `https://openreview.net/forum?id=H1aoddyvM`.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=-2FCwDKRREu`.

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. SOLAR: Deep structured representations for model-based reinforcement learning. 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 7444–7453. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/zhang19m.html`.

Denoised Predictive Imagination: An Information-theoretic approach for learning World Models (Supplementary Material)

7 Derivations

In this section, we derive equations from the Section "Denoised Predictive Imagination".

7.1 Derivation of Equation (7)

We aim to minimize the Mutual Information (MI) from the beginning to timestep t i.e. $\min I(z_{\leq t})$. To make our model action dependent, we introduce a conditional probability distribution $p(z_{t^-}, z_t|a_{t^-})$,

$$I(z_1; ...; z_t) = \mathbb{E}_{p(z_1,...,z_t)} \left[\log \frac{p(z_1, ..., z_t)}{\prod_{k=1}^{t} p(z_k)} \right]_, \tag{13}$$

$$= \mathbb{E}_{p(z_{1:t}, a_{1,t-1})} \left[\log \frac{p(z_{1:t})\, p(z_{1:t}|a_{1:t-1})}{p(z_{1:t}|a_{1:t-1}) \prod_{k=1}^{t} p(z_k)} \right]_, \tag{14}$$

$$= \mathbb{E}_{p(z_{1:t}, a_{1:t-1})} \left[\log \frac{p(z_{1:t}|a_{1:t-1})}{\prod_{k=1}^{t} p(z_k)} \right] - \mathbb{E}_{p(z_{1:t}, a_{1:t-1})} \left[\log \frac{p(z_{1:t}|a_{1:t-1})}{p(z_{1:t})} \right]_. \tag{15}$$

The first term is similar to the variational upper bound introduced in Alemi et al. [2017]. The second term is the KL-divergence between $p(z_{1:t}|a_{1:t-1})$ and $p(z_{1:t})$. Since the KL-divergence is always non-negative, the first term in the equation provides an upper bound on the MI objective we seek to optimize i.e.,

$$I(z_{1:t}) \leq \mathbb{E}_{p(z_{1:t}, a_{1:t-1})} \left[\log \frac{p(z_{1:t}|a_{1:t-1})}{\prod_{k=1}^{t} p(z_k)} \right]_. \tag{16}$$

We can write the conditional distribution $p(z_{1:t}|a_{1:t-1})$ in its autoregressive form,

$$\begin{aligned} p(z_{1:t}|a_{1:t-1}) = p(z_1, ..., z_t|a_1, ..., a_{t-1}), \\ &= p(z_t|z_{t-1}, a_{t-1}, ..., z_1, a_1)\, p(z_{t-1}, ..., z_1|a_{t-1}, ..., a_1), \\ &= p(z_t|z_{t-1}, a_{t-1}, ..., z_1, a_1)\, p(z_{t-1}|z_{t-2}, a_{t-2}, ..., z_1, a_1) \; ... \; p(z_1). \end{aligned} \tag{17}$$

To address past states and actions within the conditional distribution, we treat them as history. This history model is implemented through a Gated Recurrent Units (GRU, Cho et al. [2014]) that encapsulates these past variables into a single history variable, $h_t = \{z_{t-1}, a_{t-1}, ..., z_1, a_1\} = \{z_{t-1}, a_{t-1}, h_{t-1}\}$. Thus we can write our conditional probability in Equation (17) as,

$$p(z_{1:t}|a_{1:t-1}) = p(z_t|z_{t-1}, a_{t-1}, h_{t-1})\, p(z_{t-1}|z_{t-2}, a_{t-2}, h_{t-2}) \; ... \; p(z_1), \tag{18}$$

$$= p(z_1) \prod_{k=1}^{t-1} p(z_{k+1}|z_k, a_k, h_k). \tag{19}$$

We can substitute the conditional distribution from Equation (19) into the Upper bound in Equation (16),

$$I(z_{1:t}) \leq \mathbb{E}_{p(z_{1:t},a_{1:t-1})}\left[\log\frac{\cancel{p(z_1)}\prod_{k=1}^{t-1}p(z_{k+1}|z_k,a_k,h_k)}{\cancel{p(z_1)}\prod_{k=1}^{t-1}p(z_{k+1})}\right], \tag{20}$$

$$\leq \mathbb{E}_{p(z_{1:t},a_{1:t-1})}\left[\log\prod_{k=1}^{t-1}\frac{p(z_{k+1}|z_k,a_k,h_k)}{p(z_{k+1})}\right], \tag{21}$$

$$\leq \sum_{k=1}^{t-1}\mathbb{E}_{p(z_k,a_k)}\left[\log\frac{p(z_{k+1}|z_k,a_k,h_k)}{p(z_{k+1})}\right], \tag{22}$$

$$\leq \sum_{k=1}^{t-1} I\big(z_{k+1};z_k,a_k,h_k\big). \tag{23}$$

7.2 Derivation of Equation (9)

We aim to minimize the Mutual Information (MI) between the latent variables z_t from the beginning to time step t and the observations o_t from the environment i.e. $\min I(z_{\leq t};o_{\leq t})$, where $\cdot_{\leq t}$ is the variable from timestep 1 to t,

$$I(z_1,...,z_t;o_1,...,o_t) = \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log\frac{p(z_{1:t}|o_{1:t})}{p(z_{1:t})}\right]. \tag{24}$$

Here we introduce the conditional probability distribution $p(z_{t^-},z_t|a_{t^-})$ with the aim of removing out the denominator and including actions into our model,

$$I(z_{1:t};o_{1:t}) = \mathbb{E}_{p(z_{1:t},o_{1:t},a_{1:t-1})}\left[\log\frac{p(z_{1:t}|o_{1:t})\;p(z_{1:t}|a_{1:t-1})}{p(z_{1:t}|a_{1:t-1})\;p(z_{1:t})}\right], \tag{25}$$

$$= \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log\frac{p(z_{1:t}|o_{1:t})}{p(z_{1:t}|a_{1:t-1})}\right] - \mathbb{E}_{p(z_{1:t},a_{1:t-1})}\left[\log\frac{p(z_{1:t})}{p(z_{1:t}|a_{1:t-1})}\right], \tag{26}$$

$$= \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log\frac{p(z_{1:t}|o_{1:t})}{p(z_{1:t}|a_{1:t-1})}\right] - D_{KL}\big(p(z_{1:t})||p(z_{1:t}|a_{1:t-1})\big), \tag{27}$$

$$\leq \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log\frac{p(z_{1:t}|o_{1:t})}{p(z_{1:t}|a_{1:t-1})}\right]. \tag{28}$$

The second term is the KL-divergence between $p(z_{1:t})$ and $p(z_{1:t}|a_{1:t-1})$, which is always non-negative, leading to Equation (28) being an upper bound on our MI objective. The encodings at every timesteps depends only on that observation's timestep i.e. $p(z_{1:t}|o_{1:t}) = \prod_{k=1}^{t}p(z_k|o_k)$. Autoregressing the denominator according to Equation (19), we get,

$$I(z_{1:t};o_{1:t}) = \mathbb{E}_{p(z_{1:t},o_{1:t})}\left[\log\frac{p(z_1|o_1)\prod_{k=1}^{t-1}p(z_{k+1}|o_{k+1})}{p(z_1)\prod_{k=1}^{t-1}p(z_{k+1}|z_k,a_k,h_k)}\right], \tag{29}$$

$$= \sum_{k=1}^{t-1}\mathbb{E}_{p(z_k,o_k)}\left[\log\frac{p(z_{k+1}|o_{k+1})}{p(z_{k+1}|z_k,a_k,h_k)}\right] - D_{KL}\big(p(z_1)||p(z_1|o_1)\big). \tag{30}$$

In Equation (8), we approximate this with the transition function with variational function $q_\theta(z_{k+1}|\hat{z})$, where $\hat{z}=(z_k,a_k,h_k)$. The transition function is a neural network with parameters θ. This is the same transition function described in the Equation (9),

$$I(z_{1:t};o_{1:t}) \leq \sum_{k=1}^{t-1}\mathbb{E}_{p(z_k,o_k)}\left[\log\frac{p(z_{k+1}|o_{k+1})}{q_\theta(z_{k+1}|\hat{z})}\right] - D_{KL}\big(p(z_{k+1}|\hat{z})||q_\theta(z_{k+1}|\hat{z})\big). \tag{31}$$

As KL-divergence is non-negative, this is the upper bound on our main objective,

$$I(z_{1:t}; o_{1:t}) \leq \sum_{k=1}^{t-1} \mathbb{E}_{p(z_k, o_k)} \left[\log \frac{p(z_{k+1}|o_{k+1})}{q_\theta(z_{k+1}|z_k, a_k, h_k)} \right]_. \tag{32}$$

7.3 Derivation of Equation (10)

We aim to maximise the Mutual Information (MI) from the current timestep to the Horizon T i.e., $\max I(z_{\geq t})$; where $\geq t = \{t+1, ..., T\}$,

$$I(z_t; ...; z_T) = \mathbb{E}_{p(z_{t:T})} \left[\log \frac{p(z_{t:T})}{\prod_{k=t}^{T} p(z_k)} \right]_. \tag{33}$$

The numerator in Equation (33) can be factorised with chain rule,

$$\begin{aligned} p(z_t, ..., z_T) &= p(z_t|z_{t+1}, ..., z_T)\, p(z_{t+1}|z_{t+2}, ..., z_T) \;...\; p(z_T), &(34)\\ &= p(z_t|z_{t+1:T})\, p(z_{t+1}|z_{t+2:T}) \;...\; p(z_T), &(35)\\ &= p(z_T) \prod_{k=t}^{T-1} p(z_k|z_{k+1:T}). &(36) \end{aligned}$$

Integrating Equation (36) in Equation (33),

$$\begin{aligned} I(z_{t:T}) &= \mathbb{E}_{p(z_{t:T})} \left[\log \frac{\cancel{p(z_T)} \prod_{k=t}^{T-1} p(z_k|z_{k+1:T})}{\cancel{p(z_T)} \prod_{k=t}^{T-1} p(z_k)} \right]_, &(37)\\ &= \mathbb{E}_{p(z_{t:T})} \left[\log \prod_{k=t}^{T-1} \frac{p(z_k|z_{k+1:T})}{p(z_k)} \right]_. &(38) \end{aligned}$$

Here we incorporate conditional probability $p(z_k|z_{k+1,a_k})$ to remove $p(z_k|z_{k+1:T})$ out of our equation.

$$\begin{aligned} I(z_{t:T}) &= \mathbb{E}_{p(z_{t:T}, a_{t:T})} \left[\log \prod_{k=t}^{T-1} \frac{p(z_k|z_{k+1:T})\, p(z_k|z_{k+1}, a_k)}{p(z_k|z_{k+1}, a_k)\, p(z_k)} \right]_, &(39)\\ &= \sum_{k=t}^{T-1} \mathbb{E}_{p(z_k, a_k)} \left[\log \frac{p(z_k|z_{k+1}, a_k)}{p(z_k)} \right] + \sum_{k=t}^{T-1} D_{KL}\big(p(z_k|z_{k+1:T}) || p(z_k|z_{k+1}, a_k)\big)_, &(40)\\ &\geq \sum_{k=t}^{T-1} \mathbb{E}_{p(z_k, a_k)} \left[\log \frac{p(z_k|z_{k+1}, a_k)}{p(z_k)} \right]_, &(41)\\ &= \sum_{k=t}^{T-1} I\big(z_k; z_{k+1}, a_k\big)_. &(42) \end{aligned}$$

7.4 Derivation of Equation (12)

We aim to maximize the Mutual Information (MI) between the latent variables z_t and the observations o_t from current time step t to time-horizon T i.e. $\max I(z_{t:T}; o_{t:T})$

$$I(z_{t:T}; o_{t:T}) = \mathbb{E}_{p(z_{t:T}, o_{t:T})}\left[\log \frac{p(o_{t:T}|z_{t:T})}{p(o_{t:T})}\right], \tag{43}$$

$$= \mathbb{E}_{p(z_{t:T}, o_{t:T})}\left[\log \prod_{k=t}^{T} \frac{p(o_k|z_k)}{p(o_k)}\right]. \tag{44}$$

Introducing a tractable variational decoder with parameters ψ,

$$I(z_{t:T}; o_{t:T}) = \mathbb{E}_{p(z_{t:T}, o_{t:T})}\left[\log \prod_{k=t}^{T} \frac{p(o_k|z_k)\, r_\psi(o_k|z_k)}{r_\psi(o_k|z_k)\, p(o_k)}\right], \tag{45}$$

$$= \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)}\left[\log \frac{r_\psi(o_k|z_k)}{p(o_k)}\right] + \sum_{k=t}^{T} D_{KL}\big(p(o_k|z_k) || r_\psi(o_k|z_k)\big), \tag{46}$$

$$\geq \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)}\left[\log \frac{r_\psi(o_k|z_k)}{p(o_k)}\right], \tag{47}$$

$$= \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)}\Big[\log\, r_\psi(o_k|z_k)\Big] - \sum_{k=t}^{T} \mathbb{E}_{p(o_k)}\Big[\log\, p(o_k)\Big], \tag{48}$$

$$= \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)}\Big[\log\, r_\psi(o_k|z_k)\Big] + \sum_{k=t}^{T} H(o_k), \tag{49}$$

$$= \sum_{k=t}^{T} \mathbb{E}_{p(z_k, o_k)}\Big[\log\, r_\psi(o_k|z_k)\Big]. \tag{50}$$

The entropy term $H(o_k)$ is independent of the parameter ψ, and consequently, can be disregarded during optimization.

8 Extended Related Work

In this section, an extended related work discussion is provided.

8.1 Relation to Human Psychology

Predictive Information is maximized by the brain at a higher, more abstract level as a strategy to prevent sensory overload [Friston, 2005, Rao and Ballard, 1999]. Imagine a scenario where you're driving a vehicle and nearing a bend in the road, beyond which visibility is limited. Based on the experience of having faced congested traffic thus far (for say), you may anticipate a similar traffic configuration beyond the bend. In these instances, you mentally simulate future possibilities based on the historical experience and using the current location as a reference point. Notably, during this mental forecast, you instinctively disregard exogenous noise like vehicle's number plate, cloud formations in the sky, or roadside billboards. This subconscious omission of inconsequential details significantly influences the agent's decision-making process [Nasr et al., 2008]. While maintaining scholarly modesty, it's essential to clarify that our contribution in this paper does not constitute an ultimate solution to the challenges described. Instead, our work introduces alternative ideas, traversing similar territory and contributing fresh perspectives to the existing discourse.

Model-based Reinforcement Learning. These models simultaneously learn policy and transition dynamics, which can be used for planning, and are often sample efficient due to their ability to handle rich observations [Kaiser et al., 2020, Chua et al., 2018, Hafner et al., 2019, Ebert et al., 2018, Lowrey et al., 2019, Gelada et al., 2019, Lee et al., 2020a]. World Models Ha and Schmidhuber [2018] uses recurrent latent model to imagine future frames. Stochastic Optimal control with Latent Representations (SOLAR, Zhang et al. [2019]) model dynamics with linear-quadratic regulator. In particular, Dreamer [Hafner et al., 2020a] optimises policies via backpropogating through latent dynamics and uses recurrent state-space model for planning in latent space. These reconstruction-based methods perform effectively in standard environments. However, when exposed to environments with noise distractors, they struggle to bifurcate between information they should reconstruct and what they should disregard.

Learning Representations and RL. Recent works [Chen et al., 2020, Henaff, 2020, Tian et al., 2020] have demonstrated progress in learning representations from unlabeled data. These concepts have been integrated into reinforcement learning by works like [Laskin et al., 2020, Oord et al., 2018, Shu et al., 2020, Ma et al., 2021, Oord et al., 2018, Ma et al., 2021, Hjelm et al., 2019]. Learning invariant representations with Information-theoretic constraints have been extensively used in the literature. However, the challenge of identifying and effectively utilizing task-relevant information, which necessitates not only the preservation of predictive information but also the generation of noise-free predictions, remains largely unaddressed by most existing methods that predominantly rely on auxiliary decoders. Our concept shares similarities with PI-SAC [Lee et al., 2020b], whose objective is also centered around Predictive Information. PI-SAC aims to identify a latent representation of the current state that reduces the MI between past observations and actions $I(o_{t^-}; z_{t^-}|a_{t^-})$, while simultaneously maximising the MI between all future observations and rewards, represented as $I(o_{t^+}; r_{t^+})$. Notably, the authors of PI-SAC present this objective straightforwardly, without an underlying mathematical derivation of selection of the variables. In contrast, our method is underpinned by a solid theoretical foundation, where objectives related to latent representations and actions emerge implicitly. Furthermore, we incorporate a historical variable that circumvents the need to consider the entire trajectory by accumulating all the information in that variable, which solves the problem of considering the entire trajectory. Empirically, it has been shown in numerous previous papers that PI-SAC underperforms in scenarios involving distractors Wang et al. [2022], Liu et al. [2024], underscoring the robustness and effectiveness of our approach in such complex environments. Unlike strategies such as Dynamic Bottleneck (DB, Bai et al. [2021]) and Sequential Information Bottleneck (SIBE, You et al. [2022]), our approach not only seeks compact representations under noisy conditions, but also emphasizes on achieving noiseless future predictions and treating temporal noise along representations.

9 Implementation Details

In this section further algorithmic implementation details are discused.

9.1 Algorithm

We jointly train DPI with Soft Actor-Critic by incorporating Equation (13) as an auxiliary objective. Soft Actor-Critic (SAC) [Haarnoja et al., 2018] is an off-policy actor-critic reinforcement learning algorithm designed to optimize stochastic policies. It incorporates maximum entropy framework, ensuring a stochastic policy that seeks to balance reward maximization with entropy maximization. SAC employs a value function and two Q-functions (or critics) to reduce value overestimation. We specifically utilise the same encoder architecture as in Yarats et al. [2021]. It aims at learning the latent state representation and policy jointly.
The training algorithm for DPI with SAC is presented in Algorithm 1. E_{step} is the environment step function. φ and θ are the parameters of observation encoder and transition function respectively. They are jointly optimised. The parameters of the two Q-function and the policy π are denoted by $\{\phi_q^1, \phi_q^2\}$ and ϕ_a respectively. $\{\varphi_m, \hat{\phi_q^1}, \hat{\phi_q^2}\}$ are the parameters of the target encoder and target Q-functions respectively, which updated with an exponential moving average. α is the temperature parameter. $\lambda_Q, \lambda_\pi, \lambda_\alpha$ and λ_{DPI} are the learning rates for four different objective functions.

9.2 Model Architecture Details

Our implementation of Soft Actor-Critic [Haarnoja et al., 2018] is implemented in PyTorch and is based on the implementation of SAC-AE [Yarats et al., 2021].

Algorithm 1 Training Algorithm for SAC with DPI

```
Require: $\text{E}_{\text{step}}, \alpha, \varphi, \theta, \psi, \phi_a, \phi_q^1, \phi_q^2$, L    ▷ Environment and initial parameters.
$D \leftarrow \emptyset$    ▷ Initialize replay buffer
for each initial collection step do
    $a_t \sim \pi_{\text{random}}(\cdot|o_t)$    ▷ Sample action from a random policy
    $o_{t+1}, r_{t+1} \sim \text{E}_{\text{step}}(a_t)$    ▷ Apply action
    $D \leftarrow D \cup (o_{t+1}, a_t, r_{t+1})$    ▷ Append experience to replay buffer
end for
for every training step do
    $\{(o_t, a_t, r_t, o_{t+1})\}_{t=k}^{L+k} \sim D$    ▷ Sample minibatch of sample from buffer
    for $t = 1$ to $L$ do
        $a_t \sim \pi_{\phi_a}(a_t|o_t)$    ▷ Sample action from the policy
        $o_{t+1}, r_{t+1} \sim \text{E}_{\text{step}}(a_t)$
        $D \leftarrow D \cup (o_{t+1}, a_t, r_{t+1})$
        for each gradient step do
            $\{\phi_q^i, \varphi\} \leftarrow \{\phi_q^i, \varphi\} - \lambda_Q \nabla \mathcal{L}_Q(\phi_q^i, \varphi)$ for $i \in \{1, 2\}$    ▷ Update soft Q-functions
            $\phi_a \leftarrow \phi_a - \lambda_\pi \nabla \mathcal{L}_\pi(\phi_a)$    ▷ Update policy
            $\alpha \leftarrow \alpha - \lambda_\alpha \nabla \mathcal{L}_\alpha(\alpha)$    ▷ Adjust temperature
            $\{\varphi, \theta\} \leftarrow \{\varphi, \theta\} - \lambda_{DPI} \nabla \mathcal{L}_{DPI}(\varphi, \theta)$    ▷ Update encoder and transition model
            $\hat{\phi}_q^i \leftarrow \tau \phi_q^i + (1 - \tau)\hat{\phi}_q^i$ for $i \in \{1, 2\}$    ▷ Update target Q-function
            $\varphi_m \leftarrow \tau \varphi + (1 - \tau)\varphi_m$    ▷ Update target encoder
        end for
    end for
end for
```

9.2.1 Critic and Actor Network

For our critic, we use double Q-learning, where each Q-function is a 3-layer MLP, using ReLU activations after every layer, except the final one. Similarly, the actor is structured as a 3-layer MLP with ReLUs, designed to produce the mean and covariance values of the diagonal Gaussian. The hidden dimensions are set to 50 for actor and critic.

9.2.2 Observation Encoder and Decoder Networks

Encoder. Our encoder architecture aligns with the design proposed by Yarats et al. [2021]. The architecture starts with an initial convolutional layer featuring a 3 × 3 kernel and a stride of 2. Subsequent to this, there are three more convolutional layers, each characterized by a 3 × 3 kernel and a stride of 1, resulting in a total of four convolutional layers, which have RELU activations. The 50 dimensional output of the fully-connected layer is stabilized using layer normalization [Ba et al., 2016], then divided into mean and standard deviation. We add tanh non-linearity on the standard deviation, then perform reparameterization trick to produce encoder's representation from the given observation.

Decoder. Our decoder is structured with an initial fully connected linear layer, followed by three deconvolutional layers with a 3 × 3 kernel and with a stride of 1, and the last layer with the same kernel size and stride of 2.

9.2.3 Transition Network

Our transition model integrates representation z_t (from the encoder) and action a_t into a single encoding, denoted as za_t, of size 256 via a fully connected linear layer. This encoding is subsequently passed through three additional fully connected layers, each having the same size and all using the Exponential Linear Units (ELU) as the activation function. To incorporate temporal dependencies, the state-action encoding is merged with the past history variable h_{t-1} via a Gated Recurrent Unit (GRU) mechanism. On another hand, this state-action encoding is concatenated (z_t^{input}) and passed via a fully connected linear layer to generate the next representation mean $\mu_{z_{t+1}}$ and standard deviation

$\sigma_{z_{t+1}}$. They are then reparameterised to produce the next representation z_{t+1}. The entire procedure is comprehensively detailed in Algorithm 2.

Algorithm 2 Transition Model Pseudo-code

```
Require: $z_t, a_t, h_{t-1}$                                ▷ Representation, Action and History
$za_t \leftarrow \text{cat}(z_t, a_t)$                      ▷ Concatenate Representation and action
$za_t \leftarrow \text{ELU}(\text{fc}_1(za_t))$             ▷ Representation-action encoding
for $i = 2$ to $4$ do
    $za_t \leftarrow \text{ELU}(fc_i(za_t))$
end for
$h_t \leftarrow \text{GRU}(za_t, h_{t-1})$                  ▷ Current history variable for next representation
$z_t^{input} \leftarrow \text{cat}(za_t, h_{t-1})$          ▷ Input for encoding next representation
$\mu_{z_{t+1}} \leftarrow \text{ELU}(fc_\mu(z_t^{input}))$  ▷ Next representation mean
$\sigma_{z_{t+1}} \leftarrow \tanh(fc_\sigma(z_t^{input}))$ ▷ Next representation standard deviation
$z_{t+1} \leftarrow \mu_{z_{t+1}} + \epsilon \odot \exp(\sigma_{z_{t+1}})$  ▷ Reparameterization trick
```

9.3 Code details

Upon publication, all code will be made publicly available. Additionally, we intend to release the code for the benchmarked algorithms.

10 Hyperparameters

To ensure a fair comparison, we maintained the original hyperparameters for each method and used the code as provided by the authors. The only adjustment we made is in how background images are incorporated into the observation. The complete set of Hyperpameters essential to implement our approach are provided in the Table 2.

10.1 Sequence Length

A crucial aspect in our method is selecting the length of the time sequence. Ideally, it could span from the trajectory's start to a certain time horizon in the future. In our method, we establish that each information term can be splitted in a Markovian fashion, due to the incorporation of the history variable. For our experiments, we've chosen a time sequence length of three timesteps.

10.2 Action Repeat

Following Dreamer [Hafner et al., 2020a], we designate repeat action of 2 for each environment. We adopt the same settings for all our baselines.

10.3 Weighing Coefficients

We performed a grid search on the weighing coefficients from a range of 1 to 10^{-5}. We empirically found out that setting α_2 large makes the algorithm unstable, as the I_{CLUB} loss dominates other terms significantly. The best settings are shown in the Table 3.

Table 3: Environment and their Coefficients

Environment	**Weighing Coefficients**			
	α_1	α_2	β_1	β_2
Cheetah Run	10^{-1}	10^{-3}	10^{-2}	10^{-2}
Walker Walk	10^{-2}	10^{-4}	10^{-2}	1
Cartpole Swingup	10^{-2}	10^{-4}	10^{-1}	10^{-1}

Table 2: Hyperparameter settings and descriptions for the SAC with DPI implementation

Parameter name	Value	Description
Replay buffer capacity	2.5×10^5	Maximum number of past experiences stored for off-policy learning.
Image size	$84 \times 84 \times 3$	RGB image of size 84×84.
Batch size	32	Number of experiences sampled from the replay buffer for each update.
Discount γ	0.99	Factor by which future rewards are discounted in the Q-function.
Optimizer	Adam	Optimization algorithm used for training; Parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon_{ADAM} = 10^{-7}$.
Critic learning rate	10^{-5}	Learning rate used to update the critic's parameters.
Critic target update frequency	2	Frequency of copying weights from the critic to the target critic.
Critic Q-function soft-update rate τ_Q	0.005	Rate of soft-updating the critic's Q-function.
Critic encoder soft-update rate τ_ϕ	0.005	Rate of soft-updating the critic's encoder.
Actor learning rate	10^{-5}	Learning rate used to update the actor's parameters.
Actor update frequency	2	Frequency of actor parameter updates.
Actor log stddev bounds	[-10, 2]	Bounds on the logarithm of the actor's policy standard deviation.
Encoder learning rate	10^{-5}	Learning rate used to update the encoder's parameters.
Decoder learning rate	10^{-5}	Learning rate used to update the decoder's parameters.
Temperature learning rate	10^{-4}	Learning rate for the temperature parameter in the SAC's objective.
Init temperature	0.1	Initial temperature parameter that scales the entropy term in SAC's objective.

The coefficients are as follows, α_1: Weighing coefficient for I_{LTC}, α_2: Weighing coefficient for I_{CLUB}, β_1: Weighing coefficient for I_{Rec} and β_2: Weighing coefficient for I_{Rec}.

11 Experiments and Analysis

11.1 Videos Configuration

In this study, we slightly modified the background from what has been traditionally done in previous research. These minor alterations significantly influenced the outcomes. Our experimental conditions closely resembles that of Temporal Predictive Coding (TPC, Nguyen et al. [2021]), but we find it crucial to articulate this explicitly here.

1. Contrary to the predominant use of grayscale images in benchmarking across numerous past studies, including Denoised MDPs [Wang et al., 2022], Task Informed Abstractions (TIA, Fu et al. [2021]), Deep Bismulation for Control (DBC, Zhang et al. [2021]),Dreamer [Hafner et al., 2020a], with the notable exception of TPC (Nguyen et al. [2021]), our work deviates by employing RGB videos instead.
2. We eliminated the ground plane to fully expose the natural background in the observations.
3. In order to ensure generalizability, we leverage a large collection of videos, segregating them into distinct sets for training and testing. Specifically, we've independently sampled 100 videos each for both training and testing. These natural videos are incorporated from Kinetics 400 dataset [Kay et al., 2017] at random.

For transparent benchmarking and easy access, we will subsequently upload these videos to a cloud storage platform on publication.

11.2 Baseline Methods

DBC. We used the observation of size 84×84 and stacked 3 consecutive frames following the original work [Zhang et al., 2021]. We used the same hyperparameters mentioned in its paper.

Others. Utilizing the Recurrent State-Space Model (RSSM) as their transition model [Hafner et al., 2019], these methods follow an identical training schedule. For all the methods, we use 64×64 images and use the same parameters described in their respective papers. In order to maintain homogeneity, we used the same number of actions for all the baselines. The author's open source-code are utilised for their implementation without any changes.

11.3 Difference in Iso-Dreamer Performance

Our results demonstrate that the performance of Iso-Dreamer [Pan et al., 2022], as seen in Table 1 and Table 4, does not align with the outcomes reported in its original publication (refer to Table 1 on Page 7 in Pan et al. [2022]). This discrepancy can be attributed to our methodological approach, particularly our decision to train and test on the `video_hard` environment, as opposed to the `video_easy` environment discussed in Section 4.2 of the Iso-Dreamer paper Pan et al. [2022]. Opting for the `video_hard` environment significantly escalates the complexity of the problem. This environment presents more intricate and challenging scenarios for learning, thereby making it harder to learn noise-free representations. Additionally, to ensure consistency, we extended the training duration to 1M steps, rather than the 500K steps as mentioned in the original study. We also made sure that the rotation of the target is fixed in the Reacher Easy environment.

11.4 Results in Standard Settings

While our main focus isn't on noiseless environments, we evaluated our method against baseline approaches in such settings. We observed that Dreamer outperforms all the methods in most of the environment in these settings. As depicted in Figure 4, our method is competitive in most of the environments.

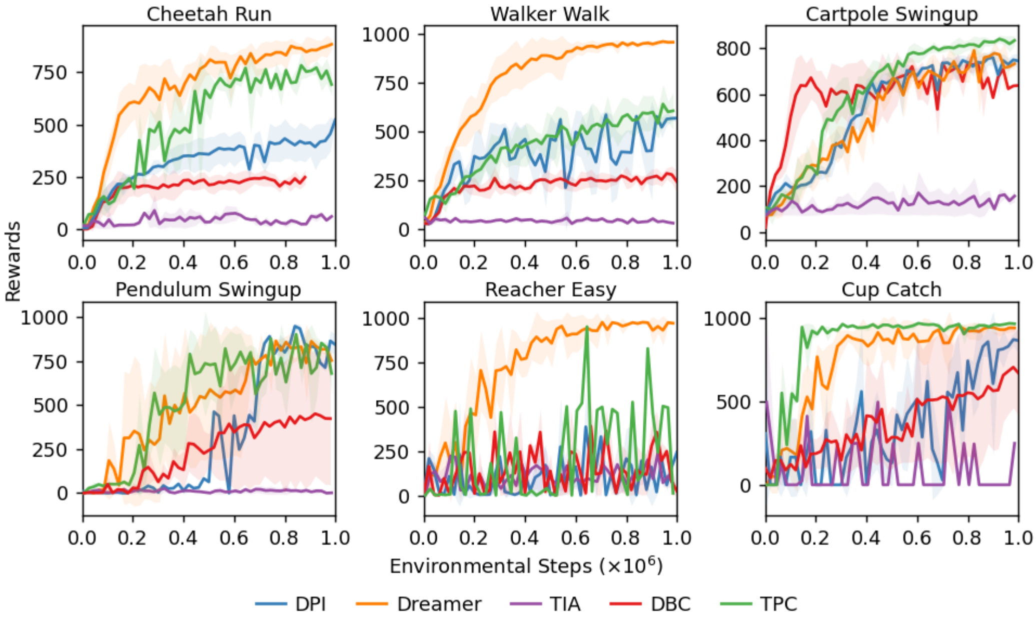


Figure 4: **Standard setting**. Performance comparison of our method (DPI) and baselines on six observation-based continuous control tasks from DMC Suite. Mean of 3 runs; shaded areas are 95% confidence intervals.

Table 4: Rewards in Random Environment Background Settings

Task	**Cheetah Run**	**Cartpole Swingup**
DBC	37 ± 3	268 ± 167
De-MDPs	118 ± 41	149 ± 38
Dreamer	118 ± 41	149 ± 38
Dreamer-V2	155 ± 103	144 ± 38
SPR	105 ± 32	201 ± 18
TIA	16 ± 7	75 ± 2
TPC	59 ± 10	132 ± 97
VSG	127 ± 79	139 ± 10
Iso-Dreamer	5 ± 2	56 ± 27
DPI (Ours)	$\mathbf{248 \pm 33}$	$\mathbf{572 \pm 110}$

The table illustrates the rewards obtained in random background settings in two environments. The best or comparable method is present in bold.

11.5 Results in Random Settings

The results for the Cheetah run task in random background settings shows the performance of DPI in comparison with nine relevant baselines. As illustrated in Figure 5, it is evident that DPI outperforms the performance of all in this setting. Denoised MDPs [Wang et al., 2022], our closest competitor here, exhibits high variance and instability, making its results less reliable and subject to fluctuation, undermining its utility in stochastic environments.

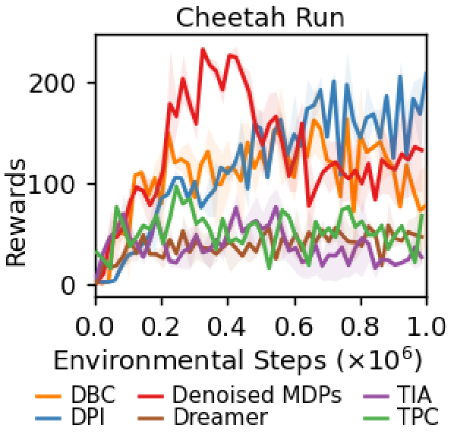


Figure 5: **Random setting**. Performance comparison of our method (DPI) and nine relevant baselines on Cheetah run environment. Mean of 3 runs; shaded areas are 95% confidence intervals.

11.6 Computational Costs

All the experiments were done on a single GPU, that required atmost 8GB memory for all the tasks. We use multiple NVIDIA GPUs for training: 4070 (DBC and DPI), 4090 (DPI and Denoised MDPs), 3090 (TPC), P500 (TIA and Dreamer). Training time required for each run heavily depends on the CPU specification too. It also heavily relies on the batch size the algorithms are trained on. Single seed of each method on average takes following time: DPI: $8 \sim 20$ hours, TIA: $15 \sim 24$ hours, Denoised MDPs: $5 \sim 8$ hours, TPC: $30 \sim 40$ hours, Dreamer: $15 \sim 24$ hours, DBC: $12 \sim 20$ hours.

12 Reconstructions

12.1 Reconstruction in the natural background setting

In our experiments, we explore the type of information encoded by different model encoders when trained in natural background settings. As depicted in Figure 6, while Dreamer (3rd row) attempts to encode both the agent and the background, DPI (2rd row) emphasizes on encoding the task-relevant agent, while the background is blurred. On the other hand, Denoised MDPs [Wang et al., 2022] also incorporate the background of other natural videos in the dataset, a consequence of overfitting on the training background noise, failing to generalise and separate the background from the agent.

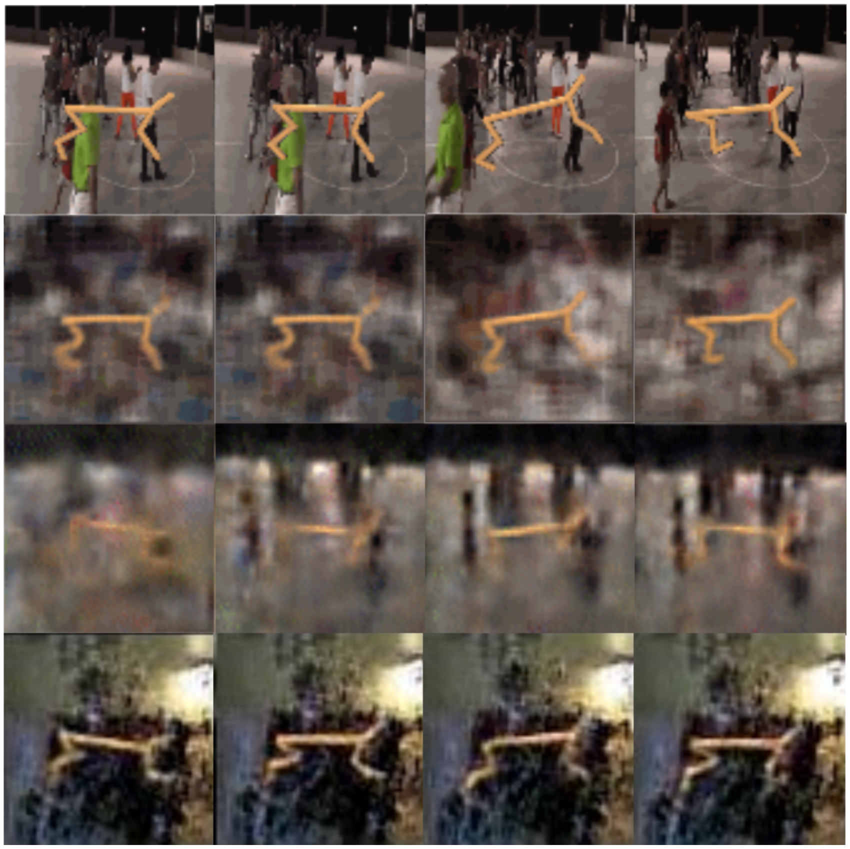


Figure 6: **Reconstruction**. Observation reconstruction of DPI versus Dreamer in the Natural background setting. First row: Ground Truth, Second row: DPI, Third row: Dreamer, Fourth Row: Denoised MDPs.

12.2 Reconstruction in blended backgrounds

We conduct experiments to investigate the challenges encountered in environments where the agent blends with their background due to similar colors. This phenomenon of color-based blending makes it difficult for the encoder to bifurcate between task-relevant features and background noise.

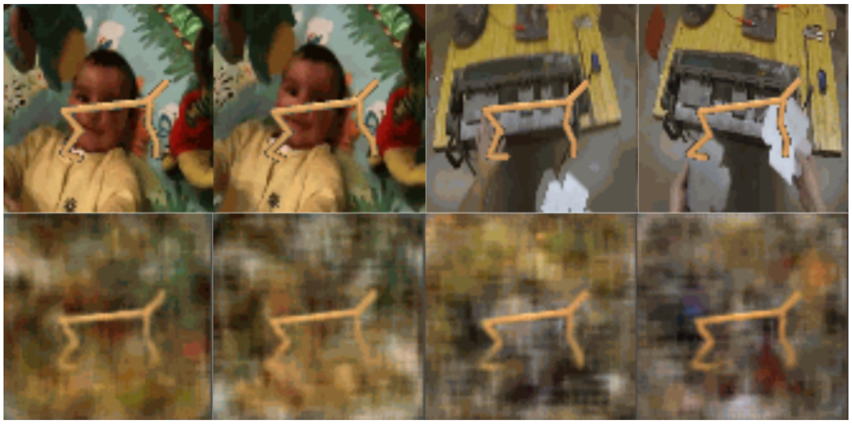


Figure 7: **Reconstruction in blended environments**. Observation reconstruction of DPI in the Natural background setting with similar color of agent and the background. First row: Ground Truth, Second row: DPI reconstruction

As illustrated in Figure 7, DPI prioritizes capturing task-relevant information and opts not to encode the background when it exhibits similar colors. In the reconstructions, the agent stands out distinctly, whereas the background appears blurred, underscoring DPI's focus on the agent over the surrounding noise.

12.3 Reconstruction of Cartpole swingup in random backgrounds

To investigate further into whether our method effectively emphasizes on relevant details, we carried out additional experiments on the Cartpole Swingup task. The findings from these experiments are shown in Figure 8.

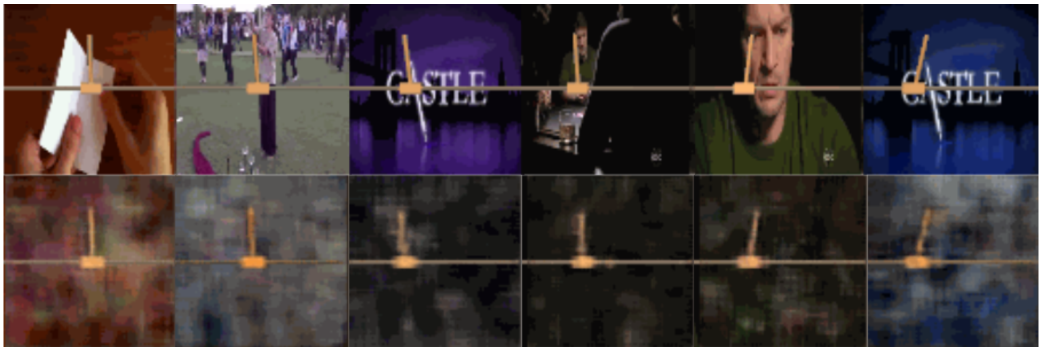


Figure 8: **Reconstruction in cartpole environment in random settings**. Observation reconstruction of DPI in the Cartpole environment in random background setting. First row: Ground Truth, Second row: DPI reconstruction

13 Ablation Analysis

In this section, we delve into an ablation study for the Cheetah Run environment, breaking down the components of the DPI model. Our experiment is conducted on various settings, each excluding distinct components in DPI (See Equation (13) for reference). Specifically, we consider:

A No latent consistency; removes I_{LTC} from $\mathcal{L}_{DPI}$ by setting $\alpha_1 = 0$.

B No upper bound minimization; removes I_{CLUB} from $\mathcal{L}_{DPI}$ by setting $\alpha_2 = 0$.

C No lower bound maximization; removes I_{NCE} from $\mathcal{L}_{DPI}$ by setting $\beta_2 = 0$.

D No reconstruction; removes I_{Rec} from $\mathcal{L}_{DPI}$ by setting $\beta_1 = 0$.

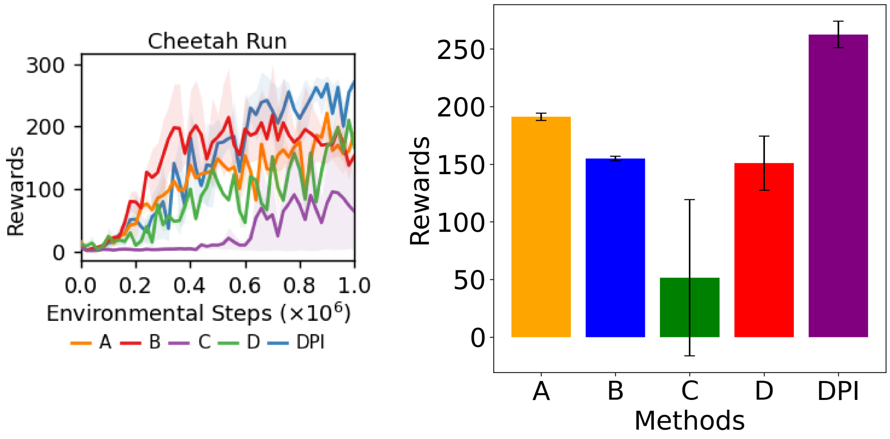


Figure 9: **Ablation Analysis**. Evaluating the impact of individual components removal on DPI's performance on Cheetah Run from DMC Suite. Mean of 3 runs; shaded areas are 95% confidence intervals.

The results of the experiments on Cheetah Run are illustrated in the Figure 9. Here we discuss the potential effects of these terms:

A No latent consistency settings eliminates the regularization of the latent representation from the transition from the observation encoder. Without this crucial regulation, the agent may deviate from its intended dynamic pathways. This deviation stems from the unmitigated noise infiltrating the predictive process due to past observations. Consequently, the agent's trajectory can diverge from its expected dynamics, underscoring the essential role of latent consistency in guiding and stabilizing the agent's behavior in alignment with its inherent dynamics. This results into a drop in performance and noise addition from the past observations into the predicted observations (Figure 10, Third row).

B No upper bound minimization setting impacts the performance and stability in natural setting. This term is responsible for finding the current state representation from the past inputs. Exclusion of this term results in added noise in the current representations, potentially leading to higher variance and reduced performance. This can be seen in the Figure 10 (Fourth Row), where the learning algorithm is not able to accurately differentiate background video from the agent and as a result induces much more noise than in original DPI's reconstruction. The results are similar to A.

C No lower bound maximization; removes I_{NCE} from $\mathcal{L}_{DPI}$ by setting $\beta_2 = 0$. This term is responsible for predictive dynamics in the latent space. Based on our findings, omitting this term most profoundly diminishes the model's performance compared to the other components. A plausible explanation might be that this term prevents the representation from collapsing by incorporating the target encoder and updating it through a moving average. This is evident in the reconstructed image shown in Figure 10 (Fourth row), where all the observations converge to a singular representation, leading to similar outputs during reconstruction. It's worth mentioning that only the agent remains and the background is entirely eliminated in this scenario. This could be attributed to I_{CLUB} taking control and effectively filtering out all the noise.

D No reconstruction; removes I_{Rec} from $\mathcal{L}_{DPI}$ by setting $\beta_1 = 0$. In a reconstruction-based approach like ours, the reconstruction loss is vital for training the model to accurately generate and replicate the complex dynamics of the observed environment. Without this component, the model's ability to accurately predict and reconstruct future states based on current observations is significantly compromised.

Concluding remarks on the Ablation study: We systematically evaluated the impact of omitting specific components within DPI, revealing their individual and collective contributions to the model's performance. Eliminating any component from our model results in a notable decline in performance, either due to the introduction of noise into our representations or a loss of the model's predictive capabilities. Our findings demonstrate that each component plays a critical role in the model's ability to accurately capture and predict the environment's dynamics.

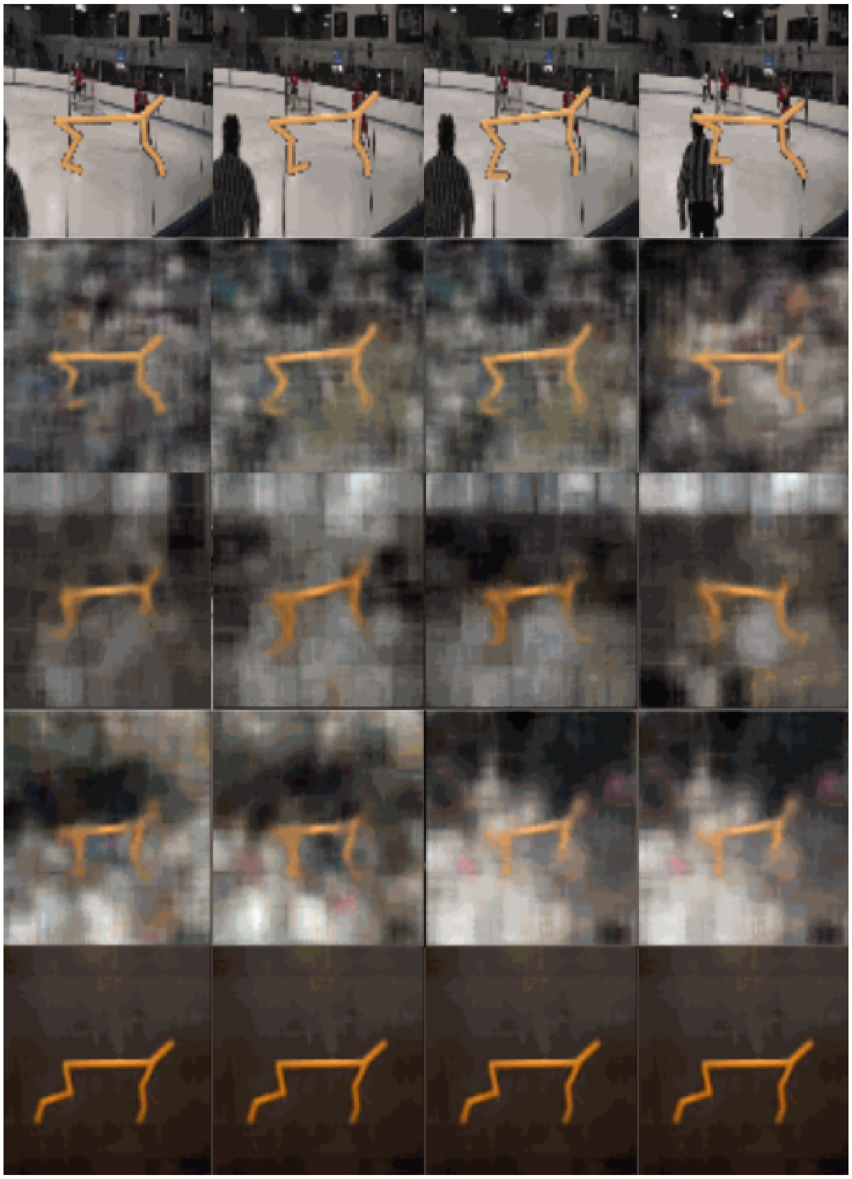


Figure 10: **Ablation Reconstruction**. Evaluating the impact of individual components removal on DPI's reconstruction on Cheetah Run from DMC Suite. First row: Ground Truth, Second row: DPI, Third row: A, Fourth row: B, Fifth row: C. We have not included D as it does not have the reconstruction.