Understanding Effectiveness of Learning Behavioral Metrics in Deep Reinforcement Learning

Anonymous authors
Paper under double-blind review

Keywords: behavioral metrics, bisimulation metrics, representation learning, evaluation

Summary

A key approach to state abstraction is approximating *behavioral metrics* (notably, bisimulation metrics) in the observation space, and embed these learned distances in the representation space. While promising for robustness to task-irrelevant noise shown in prior work, accurately estimating these metrics remains challenging, requiring various design choices that create gaps between theory and practice. Prior evaluations focus mainly on final returns, leaving the quality of learned metrics and the source of performance gains unclear. To systematically assess how metric learning works in deep RL, we evaluate five recent approaches. We unify them under isometric embedding, identify key design choices, and benchmark them with baselines across 20 state-based and 14 pixel-based tasks, spanning *250+ configurations* with diverse noise settings. Beyond final returns, we introduce the denoising factor to quantify the encoder's ability to filter distractions. To further isolate the effect of metric learning, we propose an isolated metric estimation setting, where the encoder is influenced solely by the metric loss. Our results show that metric learning improves return and denoising only marginally, as its benefits fade when key design choices, such as *layer normalization* and *self-prediction loss*, are incorporated into the baseline. We also find that commonly used benchmarks (e.g., grayscale videos, varying state-based Gaussian noise dimensions) add little difficulty, while Gaussian noise with random projection and pixel-based Gaussian noise remain challenging even for the best methods. Finally, we release an open-source, modular codebase to improve reproducibility and support future research on metric learning in deep RL.

Contribution(s)

1. We analyze five metric learning approaches under the isometric embedding framework to identify key design choices.
   Context: Metric learning methods often diverge significantly between theory and implementation.
2. We introduce the denoising factor to quantify an encoder's ability to filter distractions.
   Context: Metric learning is often motivated by denoising ability but is rarely evaluated directly, with prior work relying mainly on qualitative analysis (Zhang et al., 2020).
3. We benchmark five metric learning approaches across diverse distracting domains and find that common benchmarks add little difficulty to clean tasks, while certain noise settings remain challenging even for the best methods.
   Context: Prior work primarily uses IID Gaussian noise with varied dimensions (Ni et al., 2024) and grayscale video backgrounds (Zhang et al., 2020).
4. Through ablation studies, we identify layer normalization and self-prediction loss as key design choices across all methods.
   Context: Prior work in metric learning does not isolate the effect of self-prediction loss and only shows the benefits of normalization in specific methods (Zang et al., 2022).
5. We show that the benefits of metric learning diminish in both return and denoising factor when key design choices are incorporated into the baseline.
   Context: Prior work does not report this limitation of metric learning.

Understanding Effectiveness of Learning Behavioral Metrics in Deep Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

A key approach to state abstraction is approximating *behavioral metrics* (notably, bisimulation metrics) in the observation space, and embed these learned distances in the representation space. While promising for robustness to task-irrelevant noise shown in prior work, accurately estimating these metrics remains challenging, requiring various design choices that create gaps between theory and practice. Prior evaluations focus mainly on final returns, leaving the quality of learned metrics and the source of performance gains unclear. To systematically assess how metric learning works in deep RL, we evaluate five recent approaches. We unify them under isometric embedding, identify key design choices, and benchmark them with baselines across 20 state-based and 14 pixel-based tasks, spanning *250+ configurations* with diverse noise settings. Beyond final returns, we introduce the denoising factor to quantify the encoder's ability to filter distractions. To further isolate the effect of metric learning, we propose an isolated metric estimation setting, where the encoder is influenced solely by the metric loss.

Our results show that metric learning improves return and denoising only marginally, as its benefits fade when key design choices, such as *layer normalization* and *self-prediction loss*, are incorporated into the baseline. We also find that commonly used benchmarks (e.g., grayscale videos, varying state-based Gaussian noise dimensions) add little difficulty, while Gaussian noise with random projection and pixel-based Gaussian noise remain challenging even for the best methods. Finally, we release an open-source, modular codebase to improve reproducibility and support future research on metric learning in deep RL.[1]

1 Introduction

Real-world environments often present high-dimensional, noisy observations, posing challenges for RL. For instance, in image-based settings, task-irrelevant variations in background, lighting, and viewpoint introduce distractions. Yet, despite this observational complexity, system dynamics are typically governed by *a compact, task-relevant state*. State abstraction (Li et al., 2006; Konidaris, 2019) provides a framework for extracting such *latent representations* from raw observations, filtering out irrelevant information while preserving task-critical structure. A key principle of state abstraction is that behaviorally similar states should have similar representations. Traditionally, this is enforced through state aggregation (Singh et al., 1994; Givan et al., 2003), grouping states into discrete abstract classes based on equivalence relations. However, state aggregation lacks a measure of *how different* states are across classes and struggles with continuous representations, requiring infinitely many discrete classes.

To address this, *bisimulation metrics* (Ferns et al., 2004; 2011) and their scalable variants (Castro, 2020; Zhang et al., 2020) have been proposed to define meaningful distances between observations. These fall under the broader class of **behavioral metrics** (Castro et al., 2023), which quantify state

[1]The artifact is available at https://anonymous.4open.science/r/understanding_metric-3C44

similarity based on differences in immediate rewards and transition probabilities. By learning a metric alongside deep RL, prior work (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Chen & Pan, 2022; Zang et al., 2022) has shown progress in tackling high-dimensional, noisy tasks.

Nevertheless, the role of behavioral metric learning in deep RL (**metric learning** for short) remains unclear due to the lack of systematic evaluation. First, its effectiveness relies on accurately estimating these metrics, which is challenging in complex tasks. However, prior work primarily measures performance through *returns*, without directly assessing the quality of the learned metrics. Second, metric learning is often combined with *multiple losses* (e.g., self-prediction (Zhang et al., 2020), inverse dynamics (Kemertas & Aumentado-Armstrong, 2021)), as well as *architectural choices* (e.g., normalization, ensembles (Zang et al., 2022)), making it difficult to isolate the contribution of metric learning to performance gains. Third, most studies evaluate only *OOD generalization* in environments with *grayscale natural videos* as distractions (Zhang et al., 2020), conflating robustness with generalization. Lastly, prior evaluations (Tomar et al., 2021; Li et al., 2022) report inconsistent results for the same algorithms, raising concerns about reproducibility.

In this paper, we provide a understanding of **how metric learning works in deep RL** through a systematic evaluation of five recent approaches alongside two baselines. First, we unify these metric learning objectives under a common framework using the notion of isometric embedding, identifying key design choices for our investigation. Next, to ensure a rigorous and comprehensive evaluation, we introduce diverse distraction benchmarks by varying difficulty levels, from Gaussian noise to colored natural videos, across both state-based and pixel-based domains, tested under both ID and OOD generalization. Then, we quantify the **denoising** capability – the encoder's ability to filter out distractions. We introduce the *denoising factor*, which numerically measures how well the encoder distinguishes similar from dissimilar observations.[2] Finally, we propose an *isolated* metric estimation setting to assess metric learning's contribution to denoising, independent of other losses.

Contributions. Our main contributions are as follows:

1. **Conceptual**: We analyze five recent metric learning approaches using the isometric embedding framework to identify key design choices. In addition, we propose *denoising factor* to quantify an agent's denoising capability in distracting tasks.
2. **Comprehensive benchmarking**: We benchmark these five metric learning methods and baselines on diverse distracting variants of the DeepMind Control (DMC) suite (Tassa et al., 2018). In state-based domains, across 20 DMC tasks with 10 IID Gaussian noise settings, SimSR (Zang et al., 2022), originally evaluated in pixel-based domains, significantly outperforms other methods in both return and denoising factor. In pixel-based domains, across 14 DMC tasks with 6 image background settings, RAP (Chen & Pan, 2022) performs generally best. SAC (Haarnoja et al., 2018) and DeepMDP (Gelada et al., 2019) remain competitive baselines but are often overlooked.
3. **Reevaluating benchmark difficulty**: Surprisingly, common distracting benchmarks – varying Gaussian noise dimensions in state-based domains and grayscale videos in pixel-based domains – add little difficulty. However, Gaussian noise with random projection in state-based domains and Gaussian noise in pixel-based domains remain challenging, even for the best methods.
4. **Identifying key design choices**: We find **self-prediction loss** is crucial to SimSR's success. Notably, **(layer) normalization**, used in SimSR, consistently improves return and denoising across all metric learning methods and baselines.
5. **Marginal impact of metric learning (a bitter lesson)**: The benefits of metric learning diminish when key design choices are incorporated into the baseline.
6. **Open-source codebase**: We open-source a modular and efficient codebase to improve reproducibility in the RL community.

[2] Prior work (Zhang et al., 2020) qualitatively analyzes denoising by visualizing representations with t-SNE.

2 Background

2.1 Problem Formulation

We consider a setting where observations contain distractions and focus on a special class of Markov decision processes – exogenous block MDPs (EX-BMDPs) (Efroni et al., 2021; Islam et al., 2022). Before introducing EX-BMDPs, we first define block MDPs as a prerequisite.

Block MDPs (Du et al., 2019). A block MDP (BMDP) is a tuple $\langle \mathcal{X}, \mathcal{Z}, \mathcal{A}, q, p, R, \gamma \rangle$, where $\mathcal{X}$ is the observation space, $\mathcal{Z}$ is the latent state space, $\mathcal{A}$ is the action space, $p : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$ is latent transition function, $R : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ is (latent) reward function, and $\gamma \in [0, 1)$ is the discount factor. The emission function $q : \mathcal{Z} \to \Delta(\mathcal{X})$ generates observation $x \sim q(\cdot \mid z)$ from latent state z. Crucially, BMDP assumes the *block structure*: $\forall\, z_1, z_2 \in \mathcal{Z},\ z_1 \neq z_2 \implies \text{supp}(q(\cdot \mid z_1)) \cap \text{supp}(q(\cdot \mid z_2)) = \emptyset$. This ensures that each observation uniquely determines its latent state, enabling the existence of the *oracle encoder* $q^{-1} : \mathcal{X} \to \mathcal{Z}$ such that $q^{-1}(x) = z$ whenever $x \sim q(\cdot \mid z)$. The goal of RL in BMDP is to find a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ that maximizes the rewards: $\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(q^{-1}(x_t), a_t)\right]$. The policy only receives the observation x without access to the latent state z, the latent space $\mathcal{Z}$, or the oracle encoder q^{-1}. While the class of BMDPs is equivalent to the class of MDPs (Du et al., 2019), they capture the underlying state from a high-dimensional observation. However, BMDPs do not differentiate between task-relevant (endogenous) state and task-irrelevant (exogenous) noise in the latent space.

Exogenous BMDPs (Efroni et al., 2021). An EX-BMDP extends BMDP by factorizing a latent state into $z = (s, \xi)$, where $s \in \mathcal{S}$ is the *task-relevant state* and $\xi \in \Xi$ is the *task-irrelevant noise*, representing distraction. The latent state transition $p(s', \xi' \mid s, \xi, a)$ factorizes as $p(s' \mid s, a)p(\xi' \mid \xi)$, where the noise ξ evolves independently and does not affect the reward function. To simplify notation, we denote the reward function as as $R(s, a)$. EX-BMDPs guarantee the existence of a denoising map $D : \mathcal{Z} \to \mathcal{S}$ extracts the task-relevant state s from latent state $z \in \mathcal{Z}$. Combined with the oracle encoder in BMDPs, this enables recovery of the task-relevant state directly from observations: $s = D(q^{-1}(x))$. We define this composite function $\phi^* = D \circ q^{-1}$ as the **oracle encoder** of EX-BMDP.

2.2 Representation Learning in RL

In actor-critic methods (Konda & Tsitsiklis, 1999), representation learning is commonly used to handle complex MDPs such as EX-BMDPs. The idea is to learn an encoder that maps a raw observation to a representation, which is then shared by both actor and critic. Formally, an actor-critic algorithm employs an encoder $\phi : \mathcal{X} \to \Psi$, a (latent) actor $\pi_\theta : \Psi \to \Delta(\mathcal{A})$, and a (latent) critic $Q_\omega : \Psi \times \mathcal{A} \to \mathbb{R}$, where Ψ is the representation space. In this work, we focus on end-to-end actor-critic methods based on the soft actor-critic (**SAC**) algorithm (Haarnoja et al., 2018). These methods jointly optimize the encoder and actor-critic using the RL loss in SAC, denoted as $J_{\text{SAC}}(\phi, \theta, \omega)$.

Learning state representations solely from reward signals (i.e., RL loss) is challenging in complex tasks. To address this, various state abstraction frameworks and representation objectives have been proposed (see Ni et al. (2024) for a literature review). Among these, *model-irrelevance abstraction* (Li et al., 2006) defines two conditions for an effective encoder using *bisimulation relation* (Givan et al., 2003). The first condition, known as **reward prediction (RP)**[3], requires that the representation preserves reward information. The second condition, known as **self-prediction (ZP)**[4] (Ni et al., 2024), requires that the representation preserves latent dynamics information. Model-irrelevance abstraction thus defines compact yet informative encoders that retain sufficient information for optimal decision-making (Subramanian et al., 2022). By definition, the RP and ZP conditions hold when $\phi = \phi^*$ and $\Psi = \mathcal{S}$.[5] This implies that the oracle encoder ϕ^* serves as a model-irrelevance abstraction.

[3]Formally, in an EX-BMDP, RP condition is $\exists R_\kappa : \Psi \times \mathcal{A} \to \mathbb{R}$, s.t. $R(\phi^*(x), a) = R_\kappa(\phi(x), a), \forall x, a$.
[4]Formally, in an EX-BMDP, ZP condition is $\exists P_\nu : \Psi \times \mathcal{A} \to \Delta(\Psi)$, s.t. $P(\psi' \mid x, a) = P_\nu(\psi' \mid \phi(x), a), \forall x, a, \psi'$.
[5]In this case, $R_\kappa(s, a) = R(s, a)$ and $P_\nu(s' \mid s, a) = p(s' \mid s, a)$.

To learn a model-irrelevance abstraction, **DeepMDP** (Gelada et al., 2019) introduces **RP and ZP losses** to approximate the RP and ZP conditions, respectively. Given a data tuple (x, a, r, x'), these losses jointly optimizes the encoder ϕ, the reward model R_κ, and the latent transition model P_ν:

$$J_{\text{RP}}(\phi, \kappa) = (R_\kappa(\phi(x), a) - r)^2, \quad J_{\text{ZP}}(\phi, \nu) = -\log P_\nu(\bar{\phi}(x') \mid \phi(x), a), \tag{1}$$

where $\bar{\phi}$ detaches the encoder from gradient back-propagation. The overall objective $J_{\text{DeepMDP}}(\phi)$ for the encoder in DeepMDP combines SAC loss with RP and ZP losses (Eq. 1).

3 Conceptual Analysis on Behavioral Metrics Learning in RL

This section establishes a conceptual framework linking behavioral metrics to representations in deep RL (Sec. 3.1), and then summarizes how related work instantiates it (Sec. 3.2).

3.1 Isometric Embedding: Between Behavioral Metrics and Representation

We aim to find an encoder that maps noisy observations into a structured representation space, where distances reflect differences in rewards and transition dynamics smoothly. This representation should facilitate RL by ensuring that task-relevant variations are captured. A natural way to formalize this goal is through the concept of an *isometric embedding (isometry)*[6]:

Definition 1 (Isometric Embedding) *An encoder $\phi : \mathcal{X} \to \Psi$ is an* isometric embedding *if the distances in the original space $(\mathcal{X}, d_\mathcal{X})$ are preserved in the representation space (Ψ, d_Ψ). Formally,*

$$d_\mathcal{X}(x_1, x_2) = d_\Psi(\phi(x_1), \phi(x_2)), \quad \forall x_1, x_2 \in \mathcal{X}, \tag{2}$$

*where $d_\mathcal{X}$ is the **target metric** ("desired" metric) and d_Ψ is the **representational metric**. See Appendix Sec. B.1 for background on metric definitions.*

3.2 Design Choices in Metric Learning

Table 1: **Summary of key implementation choices for the benchmarked methods.**

Method	d_R	d_Ψ	d_T	Other Losses	Transition Model	Metric Loss	Normalization	Target Trick
SAC	—	—	—	—	—	—	—	—
DeepMDP	—	—	—	RP + ZP	Probabilistic	—	—	—
DBC	Huber	Huber	W_2 closed-form	RP + ZP	Probabilistic	MSE	—	—
RDBC	Huber	Huber	W_2 closed-form	RP + ZP	Deterministic	MSE	Max norm	—
MICo	Abs.	MICo Angular	Sample-based	—	—	Huber	—	✓
SimSR	Abs.	Cosine	Sample-based	ZP	Probabilistic Ensemble	Huber	L_2 norm	—
RAP	RAP	MICo Angular	W_2 closed-form	RP + ZP	Probabilistic	Huber	—	—

Def. 1 provides a general conceptual framework instantiated by several works in deep RL through distinct design choices. Rather than delving into theoretical implications, we focus on *practical implementations* reflected in their publicly available codebases, summarized in Table 1.

Choices of Target Metric $d_\mathcal{X}$. The target metric, which captures differences in rewards and transition dynamics, is typically formulated as (Castro et al., 2023): for $x_1, x_2 \in \mathcal{X}$,

$$d_\mathcal{X}(x_1, x_2) = c_R d_R(r_1, r_2) + c_T d_T(d_\mathcal{X})(P(\cdot \mid x_1), P(\cdot \mid x_2)), \tag{3}$$

which is inspired by bisimulation metrics (Ferns et al., 2004; 2011).[7]

- d_R, representing *immediate* state similarity, measures the closeness of sampled immediate rewards $r_1, r_2 \in \mathbb{R}$ based on x_1, x_2. Common choices include absolute difference $d_R(r_1, r_2) = |r_1 - r_2|$ ("Abs." in Table 1) (Zang et al., 2022; Castro et al., 2021), "Huber" distance[8] $d_R(r_1, r_2) = \frac{1}{2}(r_1 - r_2)^2 \mathbf{1}_{\{|r_1 - r_2| < 1\}} + \left(|r_1 - r_2| - \frac{1}{2}\right) \mathbf{1}_{\{|r_1 - r_2| \geq 1\}}$ (Huber in Table 1), or other specific forms (Chen & Pan, 2022).

[6] https://en.wikipedia.org/wiki/Isometry

[7] See Appendix Sec. B.2 for the full formal definition and discussion of the variants of bisimulation metrics.

[8] https://pytorch.org/docs/stable/generated/torch.nn.SmoothL1Loss.html

- d_T, a probabilistic measure of *long-term* state similarity, typically avoids expensive 1-Wasserstein computations in bisimulation metrics. Methods such as (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Chen & Pan, 2022) approximate transitions using a Gaussian transition model with a 2-Wasserstein metric. In contrast, Castro et al. (2021); Zang et al. (2022) rely on sample-based distance approximations.

Choices of Representational Metric d_Ψ. To approximate d_Ψ, methods employ either a Huber distance (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021) (a surrogate for L_p-distance) or an angular distance (Castro et al., 2021; Zang et al., 2022; Chen & Pan, 2022).

Metric Loss Function J_M. To approximate an isometric embedding, metric learning methods optimize this general objective:

$$J_M(\phi) = \ell(d_\Psi(\phi(x_1), \phi(x_2)) - d_\mathcal{X}(x_1, x_2)), \tag{4}$$

where ℓ can be Huber loss (Chen & Pan, 2022; Castro et al., 2021; Zang et al., 2022), or mean square error (MSE) (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021).

Self-prediction (ZP). As discussed, approximating $d_\mathcal{X}$ often requires a transition model, methods adopt distinct approaches: probabilistic models (Zhang et al., 2020; Castro et al., 2021), ensembles of probabilistic models (Zang et al., 2022), and deterministic models (Kemertas & Aumentado-Armstrong, 2021). MICo, in contrast, employs a sample-based target metric that is free of ZP.

Normalization. We discuss normalization in the representation space Ψ. RDBC (Kemertas & Aumentado-Armstrong, 2021) employs max normalization to enforce boundedness, leveraging prior knowledge of target metric constraints. SimSR (Zang et al., 2022) applies L_2 normalization to enforce unit-length representations. LayerNorm (Ba et al., 2016) is widely used in pixel-based encoder across all methods except SimSR. In the broader context of deep RL, normalization has been extensively studied for its benefits in stabilizing training and generalization (Li et al., 2023; Fujimoto et al., 2023; Elsayed et al., 2024; Gallici et al., 2024).

Target trick. MICo employs a target network $\bar{\phi}$ for encoding one observation in d_Ψ when approximating $d_\mathcal{X}$ to ensure learning stability. See Castro et al. (2021, Appendix C.2) for further details.

3.3 Candidate Methods

We present the design choices of methods to be benchmarked in our work in Table 1. Note that RDBC (Kemertas & Aumentado-Armstrong, 2021) incorporates additional components – such as intrinsic rewards and inverse dynamics – that enhance performance. However, since these elements are orthogonal to our study, they are not included in our implementation. For a fair comparison, our experiments employ a probabilistic transition model for all methods that require one.

4 Study Design: Does Metric Learning Help with Denoising?

The "denoising capability" of behavioral metric learning is often cited as a motivation in prior work (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Chen & Pan, 2022; Zang et al., 2022). However, most studies evaluate this *indirectly* by (1) combining metric learning with RL, (2) training only on grayscale natural video backgrounds, (3) testing on unseen videos in training, and (4) evaluating solely through return performance. This leaves a gap between motivation and actual denoising assessment.

This section bridges that gap with a systematic study design. First, we introduce a diverse range of noise settings from IID Gaussian noise and random projections to natural video backgrounds (Sec. 4.1), enabling an analysis of how noise difficulty impacts metric learning. Second, we separate the noise distributions during training and testing to examine denoising under both ID and OOD

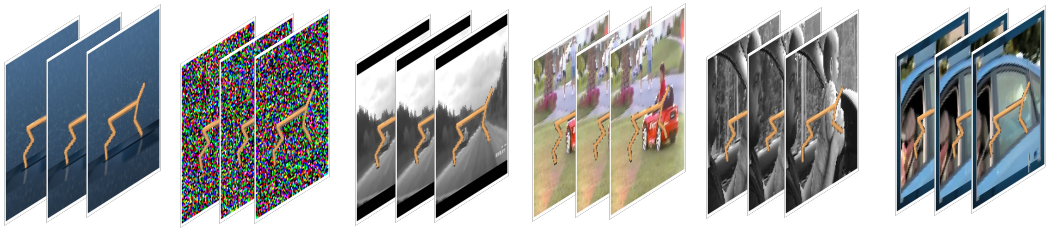

Figure 1: Examples of different noise settings in pixel-based domains (three consecutive timesteps each). **From left to right**: original clean image, IID Gaussian image, grayscale natural image, colored natural image, grayscale natural video, colored natural video. Three background images are different for video settings.

generalization settings (Sec. 4.2). Third, we introduce a direct evaluation measure, the *denoising factor* (Sec. 4.3). Finally, to disentangle metric learning from RL, we propose the *isolated metric estimation* setting, where metric learning affects only the encoder, not the RL agent (Sec. 4.4).

4.1 Noise Settings

We describe our noise settings using the EX-BMDP framework (Sec. 2.1), where observations follow $x \sim q(\cdot \mid z)$ with $z = (s, \xi)$.

IID Gaussian Noise. The task-irrelevant noise ξ_t is sampled independently at each timestep from an m-dimensional isotropic Gaussian, $\xi_t \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. For state-based domains, the observation is exactly the latent state, i.e., $x_t = z_t$ with q as the identity mapping. We adjust the noise dimension m or noise std σ to modulate difficulty, whereas prior work (Kemertas & Aumentado-Armstrong, 2021; Ni et al., 2024) only varies m with a small σ. For pixel-based domains, noise is applied per pixel in the background and overlaid by the robot foreground's pixels, with q as a rendering function.

IID Gaussian Noise with Random Projection. This setting applies only to state-based domains where $s \in \mathbb{R}^n$. Before interaction with the MDP, we construct a full-rank square matrix $\mathbf{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ with entries sampled as $A_{ij} \sim \mathcal{N}(\mu_A, \sigma_A^2)$. At each time step, we generate m-dimensional IID Gaussian noise $\xi_t \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and then apply a linear projection to obtain observation $x_t = \mathbf{A}z_t$ where $z_t = (s_t, \xi_t)$. Since $\mathbf{A}$ is full rank, s_t can be recovered from x_t using $\mathbf{A}^{-1}$. This setting is more challenging than IID Gaussian noise, as it linearly entangles s_t and ξ_t, with q as the linear projection.[9]

Natural Images. This setting applies only to pixel-based domains, replacing the clean background with a randomly selected natural image. As in the original environment, the background remains fixed during training. Images can be *grayscale* or *colored*, introducing different levels of visual complexity. In EX-BMDP notation, ξ_t is stationary and q is a rendering function.

Natural Videos. This setting also applies only to pixel-based domains, replacing the clean background with with natural video. The underlying noise $\xi_t \in \mathbb{N}$, representing the *frame index*, follows the update rule $\xi_t = (\xi_{t-1} + 1) \mod N$, where N is the total number of frames. These videos can be *grayscale* or *colored*, with the grayscale setting widely used in metric learning (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Zang et al., 2022; Chen & Pan, 2022).

4.2 Denoising Involves ID and OOD Generalization

In this work, "*denoising*" refers to a form of generalization that removes task-irrelevant noise from observations, enabling generalization to tasks with unseen noise. The evaluation settings differ based on whether the *noise distribution* remains unchanged or shifts between training and testing.

[9] Voelcker et al. (2024) similarly projects observations using a random binary matrix in discrete domains.

In-distribution (ID) Generalization Evaluation. The training and testing environments (EX-BMDPs) are identical, meaning the same noise distribution is applied in both phases. For example, IID Gaussian noise remains unchanged throughout training and testing.

Out-of-distribution (OOD) Generalization Evaluation. The training and testing EX-BMDPs share the same task-relevant parts (i.e., $p(s' \mid s, a)$, $p(s_0)$, $R(s, a)$) but differ in noise distributions (i.e., $p(\xi' \mid \xi)$, $p(\xi_0)$). For instance, natural videos from a training dataset are employed during training, while videos from a distinct test dataset are used during evaluation. This evaluation setup is widely used in metric learning (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Zang et al., 2022; Chen & Pan, 2022).

4.3 Quantifying Denoising via the Denoising Factor

We introduce the *denoising factor* (DF), a measure that quantifies an encoder ϕ's ability to filter out irrelevant details while retaining essential information.[10] It also provides insight into how the behavioral metrics are approximated, given that exact behavioral metrics are nearly inaccessible via fixed-point iteration in high-dimensional state or action spaces. To compute DF, we define a *positive score* and a *negative score* for an encoder ϕ. Inspired by triplet loss (Schroff et al., 2015) in contrastive learning, we compute these scores by selecting an observation x as an *anchor*, then constructing a *positive example* x^+ similar to x, and a *negative example* x^- dissimilar to x.

Definition 2 (Positive score) *The positive score of an encoder ϕ w.r.t. the metric d_Φ measures the average representational distance between anchors and their positive examples:*

$$\mathrm{Pos}_{d_\Phi}(\phi) := \mathbb{E}_{x\sim\rho_\pi(x),\xi^+\sim\rho(\xi^+),x^+\sim q(\cdot|\phi^*(x),\xi^+)}\big[d_\Phi(\phi(x),\phi(x^+))\big], \tag{5}$$

where $\rho_\pi(x)$ is the stationary state distribution under the policy π and $\rho(\xi^+)$ is a stationary noise distribution. The sampling $x^+ \sim q(\cdot \mid \phi^*(x), \xi^+)$ ensures that x^+ shares the same task-relevant state $s = \phi^*(x)$ but has different noise ξ^+.

In the temporally-independent noise setting, $\rho(\xi^+)$ matches the noise transition; in the natural-video setting, $\rho(\xi^+)$ is a uniform distribution over frame indices $\{0, 1, \ldots, N-1\}$.

Definition 3 (Negative score) *The negative score of an encoder ϕ w.r.t. the metric d_Φ measures the average representational distance between anchors and their negative examples:*

$$\mathrm{Neg}_{d_\Phi}(\phi) := \mathbb{E}_{x,x^-\overset{\text{iid}}{\sim}\rho_\pi}\big[d_\Phi(\phi(x),\phi(x^-))\big], \tag{6}$$

where x, x^- are IID samples from ρ_π.

Definition 4 (Denoising factor (DF)) *The denoising factor of an encoder ϕ w.r.t. the metric d_Φ is defined as the normalized difference between the negative and positive scores. As a result, a higher DF indicates better denoising ability.*

$$\mathrm{DF}_{d_\Phi}(\phi) := \frac{\mathrm{Neg}_{d_\Phi}(\phi) - \mathrm{Pos}_{d_\Phi}(\phi)}{\mathrm{Neg}_{d_\Phi}(\phi) + \mathrm{Pos}_{d_\Phi}(\phi)} \in [-1, 1]. \tag{7}$$

4.4 Decoupling Behavioral Metric Learning from RL

In many behavioral metric learning methods, the encoder ϕ is optimized via a combination of losses: the RL loss (e.g., $J_{\mathrm{SAC}}(\phi)$), the reward-prediction loss $J_{\mathrm{RP}}(\phi)$, the self-prediction loss $J_{\mathrm{ZP}}(\phi)$ (Eq. 1), and a metric loss $J_{\mathrm{M}}(\phi)$ (3). This coupling makes it difficult to isolate the direct impact of metric learning on representation quality. Moreover, denoising factor (DF, Def. 4) depends on both the encoder and the data distribution induced by RL agent. Policies that frequently revisit similar task-relevant states under varying noise may inflate DF, making it an unreliable measure of denoising. Due to the above reasons, we propose to evaluate behavioral metric learning algorithms in an *isolated metric estimation* setting.

[10] While the oracle encoder ϕ^* achieves perfect denoising, direct comparison is impossible as ϕ lacks access to $\mathcal{S}$.

Isolated Metric Estimation Setting. To isolate the effect of metric learning, we introduce an isolated *metric encoder* $\tilde{\phi}$ that is optimized solely via the metric loss $J_{\mathrm{M}}(\tilde{\phi})$, while the *agent encoder* ϕ is updated using the remaining training objectives (*e.g.*, $J_{\mathrm{SAC}}(\phi)$ or $J_{\mathrm{DeepMDP}}(\phi)$). In our experiments, regardless of the metric learning method, a SAC agent interacts with the environment and collect data for learning the metrics. This allows for a fair comparison of $\mathrm{DF}_{d_\Phi}(\tilde{\phi})$ across different metric learning methods. For methods that rely on self-prediction loss (Zhang et al., 2020; Kemertas & Aumentado-Armstrong, 2021; Zang et al., 2022), we learn an *isolated transition model* using $\tilde{\phi}$ while preventing gradient backpropagation to $\tilde{\phi}$ to ensure isolation.

5 Experiments

Experiment Organization. We first conduct a comprehensive evaluation of all the methods (Table 1) across **20 state-based** DeepMind Control (DMC) (Tassa et al., 2018; Tunyasuvunakool et al., 2020) tasks (listed in Table 5) and **14 pixel-based** DMC tasks (listed in Table 6). Evaluations are performed under various noise settings using ID generalization. This study covers a significantly larger task set than prior works. Our results (Sec. 5.1) provide a broad assessment of agent performance and task difficulty, as reflected by return variations. Based on these findings, we select a subset of representative tasks for fine-grained analysis (Sec. 5.2) to examine the impact of key design choices (Sec. 3.2). We further investigate the isolated metric evaluation setting (Sec. 4.4) in Sec. 5.3, and assess OOD generalization following prior work in Sec. 5.4.

Evaluation Protocol. We report the *mean episodic reward* rather than the IQM (Agarwal et al., 2021) to avoid ignoring tasks that are too easy or too challenging. For each run, the reported mean episodic reward, bounded within $[0, 1000]$ for all tasks, is the average of 10 evaluation points within a 1.95M-2.05M step window and aggregated over seeds. Figures and tables display 95% confidence intervals over tasks. For state-based environments, we use 12 random seeds per *configuration*, where a configuration is defined as a *(task, noise setting)* pair. For pixel-based experiments, we use 5 random seeds per configuration.

Approximation of Denoising Factor (Eq. 7). All observations collected in the evaluation stage are regarded as *anchors*. We sample 16 positive samples and negative samples for each anchor using the strategy shown in Sec. 4.3. For consistency, we report $\mathrm{DF}_{\|\cdot\|_2}(\phi)$.

5.1 Benchmarking Methods on Various Noise Settings

Settings. For state-based DMC tasks, we apply IID Gaussian noise, varying either (a) standard deviations $\sigma \in \{0.2, 1.0, 2.0, 4.0, 8.0\}$ (with a fixed noise dimension $m = 32$), or (b) noise dimensions $m \in \{2, 16, 32, 64, 128\}$ (with a fixed standard deviation $\sigma = 1.0$). For pixel-based DMC tasks, evaluation is conducted under **6 image background settings**: (1) clean background (the original pixel-based DMC setting), (2) grayscale natural images, (3) colored natural images, (4) grayscale natural videos, (5) colored natural videos, and (6) IID Gaussian noise (with $\sigma = 1.0$). ID generalization evaluation is conducted in this subsection. The aggregated reward and DF for settings (a), (b), and (1)-(6) are shown in Fig. 2 and Fig. 3, respectively. Per-task results are listed in Appendix Sec. E.

Implementation Details. For state-based tasks, the encoder is a three-layered MLP, as used by SAC (Haarnoja et al., 2018) and RDBC (Kemertas & Aumentado-Armstrong, 2021). For pixel-based tasks, the encoder is a CNN followed by LayerNorm (Ba et al., 2016), as used by SAC-AE (Yarats et al., 2021b). All the compared methods are implemented based on SAC. For a fair comparison, we adopt an identical probabilistic latent transition models and reward models used in DBC and RDBC, although some methods, such as SimSR (Zang et al., 2022), propose using an ensemble of latent transition models. We follow the exact hyperparameters specific to each metric learning method. Please see Appendix Sec. D for further details.

Benchmarking Findings. We summarize the key findings from Fig. 2 and Fig. 3.

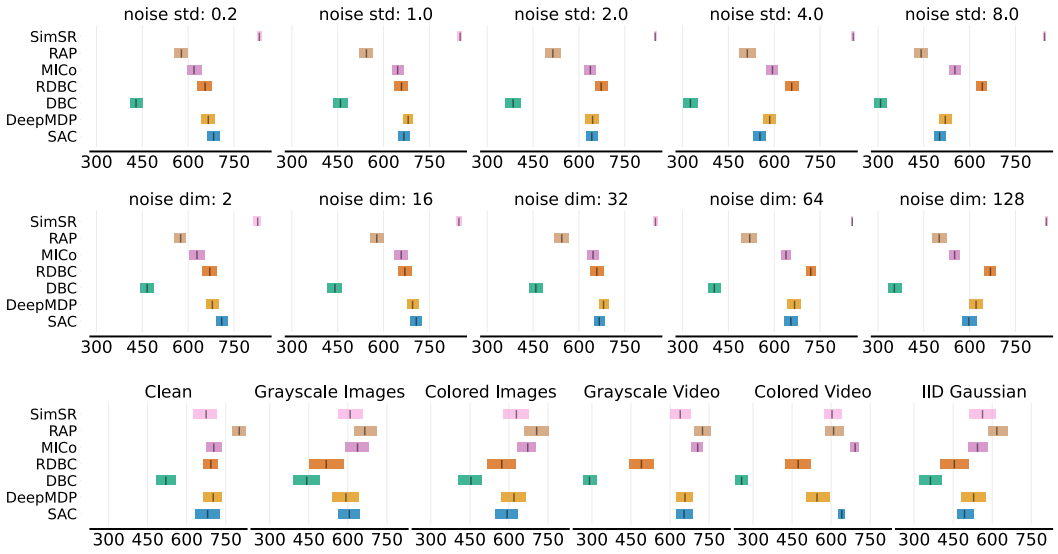

Figure 2: **Benchmarking results: performance of seven methods across diverse noise settings**, aggregating episodic rewards from **20 state-based** (first two rows) and **14 pixel-based** tasks (last row). "Noise std" denotes the IID Gaussian noise's standard deviation σ, while "noise dim" denotes its dimension m. Bars show 95% CI.

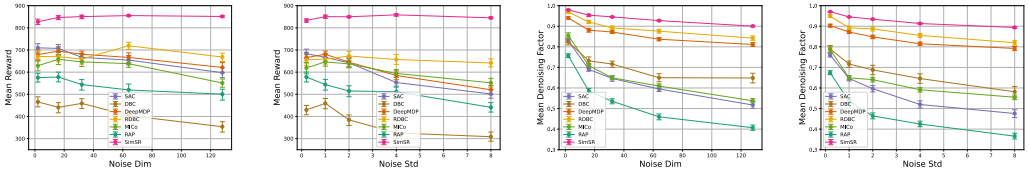

Figure 3: **Benchmarking results: reward (left) and denoising factor (right)** of seven methods to IID Gaussian noise dimension (Noise Dim) and standard deviation (Noise Std). Each point is aggregated by 20 state-based tasks in Table 5.

- **SimSR** consistently achieves the highest performance in most state-based tasks, excelling in both return and denoising factor. **RAP** performs best in most pixel-based tasks but suffers a moderate performance drop in state-based tasks. Interestingly, both SimSR and RAP were evaluated only in pixel-based domains in their papers, making our state-based findings novel.
- **SAC and DeepMDP**, as fundamental metric learning baselines, deliver decent performance on both pixel-based and state-based tasks. However, they are often overlooked in prior work.
- In state-based domains, all methods are generally more robust to increased noise *dimensions* (when $\sigma = 1.0$), as commonly used in prior work, than to increased *standard deviation* (when $m = 32$), as shown in Fig. 3.
- In pixel-based domains, *grayscale natural video*, widely used in prior work, is not significantly harder than clean background setting (e.g., for SAC and DeepMDP). Surprisingly, the *IID Gaussian noise* setting is the most challenging, warranting further study.
- Different algorithms excel in different tasks (Table 5, Table 6), e.g., RAP in reacher/easy, MICo in point_mass/easy (Fig. 23). Broad task coverage is essential to ensure generalizable insights.
- Adding objectives trades-off computation efficiency. As shown in Table 2, the time cost of optimizing of a metric loss is close to optimizing a ZP loss by comparing MICo with DeepMDP.

5.2 What Matters in Metric and Representation Learning?

To identify key factors in metric learning in state-based domains, we conduct case studies on the design choices outlined in Sec. 3.2. We select *three medium-to-hard DMC tasks*, finger/turn_easy, figure/turn_hard, quadruped/run for detailed analysis. First, a notable difference between our default encoder implementations for state-based and pixel-based tasks is the inclusion of normalization, which may significantly impact benchmarking outcomes. SimSR, the best-performing algorithm in state-based environments, employs L_2 normalization in the representation space and discusses its

Table 2: **Relative time spent on model updates** on NVIDIA L40S GPUs under the same task (walker/walk, with $\mathcal{S} = \mathbb{R}^{24}$ and $\Xi = \mathbb{R}^{32}$). Values represent the multiple of SAC's updating time. Key hyperparameters affecting the speed are set identically to Table 7.

	SAC	DeepMDP	DBC	RDBC	MICo	RAP	SimSR
Pixel-based	1.00	1.44	2.03	2.12	1.53	2.20	1.75
State-based	1.00	1.42	1.76	1.95	1.39	2.08	1.68

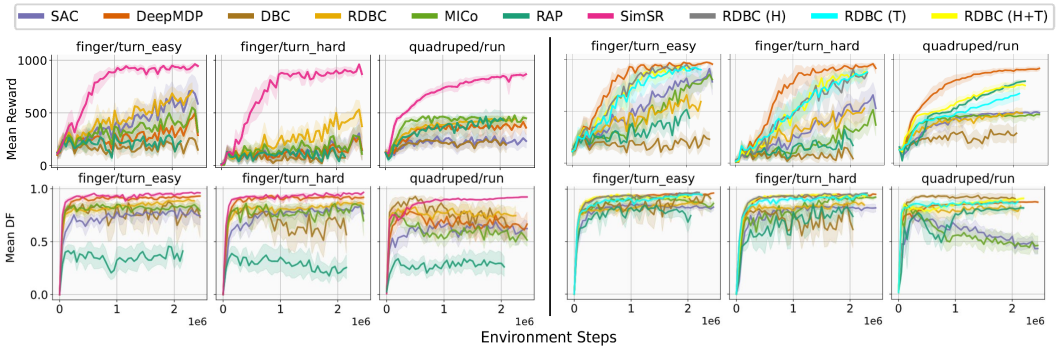

Figure 4: Case study on three DMC state-based tasks (IID Gaussian noise, noise dim=32, noise std=8.0), examining the effects of **including LayerNorm** (left vs. right three columns), **applying the target trick** (RDBC (T)), and **using Huber loss** (RDBC (H)) instead of MSE as the metric loss.

effectiveness (Zang et al., 2022). This inspires us to examine whether normalization benefits *other* metric learning methods. Second, several techniques used by the best-performing methods merit further analysis. For instance, SimSR, RAP, and MICo (which excels in *colored natural video* settings) utilize Huber metric loss instead of MSE, while MICo incorporates the target trick (Sec. 3.2). Third, we investigate the performance of methods *with LayerNorm* in a challenging setting: *IID Gaussian noise with random projection* (Sec. 4.1) with $\sigma \in \{0.2, 2.0, 4.0, 8.0\}$ (with a fixed noise dimension $m = 32$), shown in Fig. 6. Important findings from Fig. 4 to Fig. 6 is as follows:

- **Most methods benefit from LayerNorm in the representation space**, improving both reward and DF (Fig. 4). Notably, **DeepMDP with LayerNorm performs comparably to SimSR**.[11] For RDBC, using Huber loss for the metric and using the target trick enhance performance (Fig. 4).[12]
- **ZP loss is crucial for SimSR's success** in noisy state-based tasks (Fig. 5).
- A significant performance drop occurs for all agents when increasing the noise standard deviation in **IID Gaussian with random projection** setting (Fig. 6), even with normalization applied. Nevertheless, DeepMDP and SimSR remain relatively robust to the noise.

5.3 Isolated Metric Evaluation Setting: Does Learned Metrics Denoise?

We further analyze the proposed setting in Sec. 4.4 in both state-based (shown in Fig. 7) and pixel-based settings (shown from Fig. 25 to Fig. 29). We observe that:

- Generally, metrics learned in isolation achieve some denoising but underperform compared to those co-learned with ZP and critic losses or even with ZP and critic losses alone in DeepMDP (first two rows of Fig. 7).
- LayerNorm also works well with isolated metric estimation (last two rows of Fig. 7).
- MICo's DF remains relatively low, which aligns with its theoretical property that the metric for positive samples is non-zero (Fig. 7), as MICo does not enforce zero self-distance.

[11] Our additional experiments reveal that removing LayerNorm in pixel-based environments causes a substantial performance drop across all methods, highlighting the critical role of normalization.

[12] We observe that runs with superior design choices exhibit much more stable representation norms and gradient norms for both the critic loss and metric loss. Thus, the performance gains shown in Fig. 4 are likely due to improved numerical stability in extrapolating the metrics and Q values during generalization.

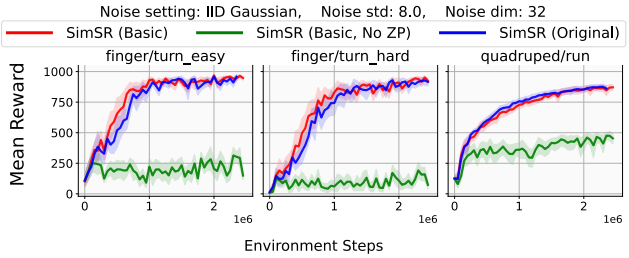

Figure 5: **Ablation study on ZP loss on SimSR.** "SimSR (Original)" is the configuration benchmarked in Sec. 5.1, where ZP is integral to the metric estimation. Therefore, we resort to "SimSR (Basic)" setting (Theorem 2, Zang et al. (2022)), where the ZP component is independent from the metric estimation, and "SimSR (Basic, No ZP)" is the setting that ZP is detached from SimSR (Basic). This ablation highlights the impact of detaching ZP on the overall performance.

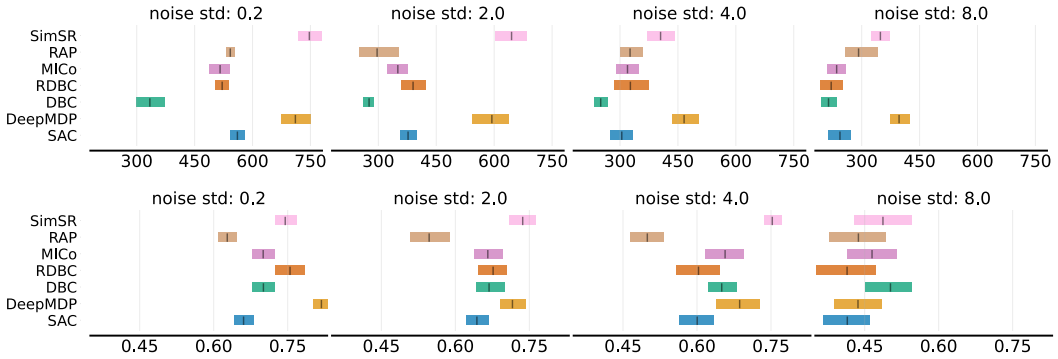

Figure 6: Aggregated reward (top row) and DF (bottom row) of seven agents on various **IID Gaussian with random projection** settings in the 3 selected state-based tasks.

5.4 OOD Generalization

While prior work has focused on OOD generalization in pixel-based settings, we extend this analysis by evaluating all **14 pixel-based tasks**. The "grayscale video" setting (and similarly for other settings) in Fig. 8 denotes using grayscale videos for both training and evaluation, with distinct video samples in each phase. Takeaways in Fig. 8 and Fig. 9 are as follows:

- Comparing Fig. 8 (OOD) with Fig. 2 (ID), all methods struggle to generalize in both grayscale and colored image settings. Unlike video backgrounds, which provide temporal variation, static images lack diverse cues, making adaptation to unseen backgrounds more challenging.
- Even with OOD generalization evaluation, SAC and DeepMDP remain competitive baselines (Fig. 8).
- OOD generalization is more challenging in the *colored* video setting than in the *grayscale* video setting (Fig. 9). Surprisingly, even baselines like SAC exhibit a low reward gap, questioning the necessity of incorporating a metric loss in the widely-used grayscale setting.

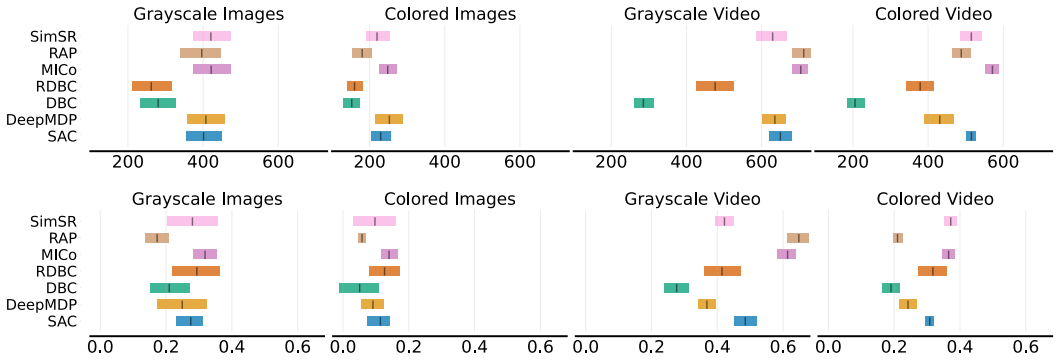

Figure 8: Aggregated reward (top row) and DF (bottom row) of seven agents on various noise settings in 14 pixel-based tasks in Table 6 with **OOD generalization** evaluation.

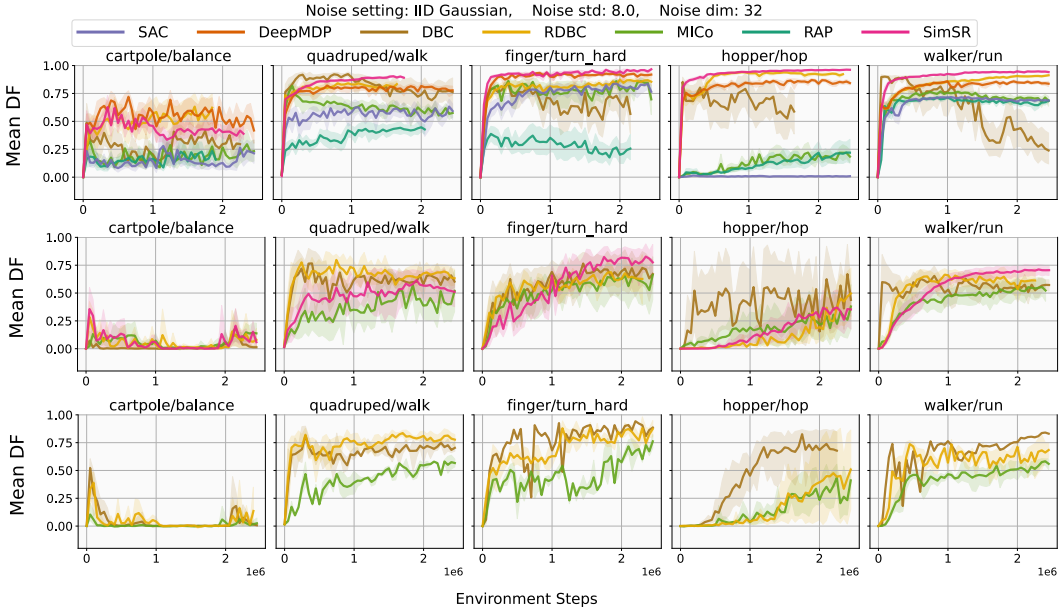

Figure 7: **Top row**: DF for the agent encoder ϕ (co-trained with RL in Sec. 5.1) without LayerNorm. **Mid row**: DF for the isolated metric encoder $\tilde{\phi}$ without LayerNorm. **Bottom row**: DF for $\tilde{\phi}$ with LayerNorm. All use ID generalization evaluation.

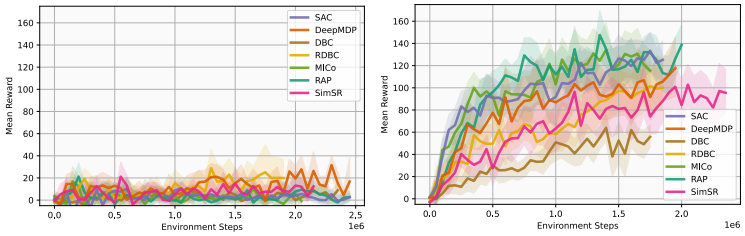

Figure 9: **Reward gap** (performance in ID evaluation minus OOD evaluation) in the grayscale video setting (left) and the colored video setting (right), aggregated on 14 pixel-based tasks in Table 6.

6 Conclusion

Takeaways. Based on our empirical studies on behavioral metric learning in deep RL, we highlight the following key insights for RL researchers:

1. To gain a clearer understanding of RL algorithms, initial evaluations should be conducted on simple, controlled environments (e.g., varying Gaussian noise std, pixel-based Gaussian noise).
2. Claims and motivations for metric learning should be supported by rigorous evaluation, including measures such as the denoising factor and comparisons between ID and OOD generalization.
3. Self-prediction loss and LayerNorm are critical design choices that significantly impact metric and representation learning.
4. The benefits of metric learning diminish when key design choices, such as self-prediction loss and LayerNorm, are integrated into SAC. This calls for a deeper investigation into when and how metric learning provides unique advantages beyond these existing techniques.

Future work. Our study focuses on continuous control; future work should explore discrete domains and real-world tasks. The relationship between denoising and return performance remains unclear, requiring further investigation. Additionally, improved benchmarks are needed to better isolate the effects of metric learning.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

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

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10069–10076, 2020.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. Mico: Improved representations via sampling-based state similarity for markov decision processes. *Advances in Neural Information Processing Systems*, 34:30113–30126, 2021.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. A kernel perspective on behavioural metrics for markov decision processes. *arXiv preprint arXiv:2310.19804*, 2023.

Jianda Chen and Sinno Pan. Learning representations via a robust behavioral metric for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36654–36666, 2022.

Edmund M Clarke. Model checking. In *Foundations of Software Technology and Theoretical Computer Science: 17th Conference Kharagpur, India, December 18–20, 1997 Proceedings 17*, pp. 54–56. Springer, 1997.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provable rl with exogenous distractors via multistep inverse dynamics. *arXiv preprint arXiv:2110.08847*, 2021.

Mohamed Elsayed, Gautham Vasan, and A Rupam Mahmood. Streaming deep reinforcement learning finally works. *arXiv preprint arXiv:2410.14606*, 2024.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 162–169, 2004.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

Scott Fujimoto, Wei-Di Chang, Edward Smith, Shixiang Shane Gu, Doina Precup, and David Meger. For sale: State-action representation learning for deep reinforcement learning. *Advances in neural information processing systems*, 36:61573–61624, 2023.

Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob Nicolaus Foerster, and Mario Martin. Simplifying deep temporal difference learning. *arXiv preprint arXiv:2407.04811*, 2024.

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*, pp. 2170–2179. PMLR, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial intelligence*, 147(1-2):163–223, 2003.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Didolkar, Dipendra Misra, Xin Li, Harm Van Seijen, Remi Tachet des Combes, et al. Agent-controller representations: Principled offline rl with rich exogenous information. *arXiv preprint arXiv:2211.00164*, 2022.

Mete Kemertas and Tristan Aumentado-Armstrong. Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems*, 34:4764–4777, 2021.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

George Konidaris. On the necessity of abstraction. *Current opinion in behavioral sciences*, 29:1–7, 2019.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

Lu Li, Jiafei Lyu, Guozheng Ma, Zilin Wang, Zhenjie Yang, Xiu Li, and Zhiheng Li. Normalization enhances generalization in visual reinforcement learning. *arXiv preprint arXiv:2306.00656*, 2023.

Xiang Li, Jinghuan Shang, Srijan Das, and Michael Ryoo. Does self-supervised learning really improve reinforcement learning from pixels? *Advances in Neural Information Processing Systems*, 35:30865–30881, 2022.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive rl. *arXiv preprint arXiv:2401.08898*, 2024.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Satinder Singh, Tommi Jaakkola, and Michael Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, 7, 1994.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

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.

Manan Tomar, Utkarsh A Mishra, Amy Zhang, and Matthew E Taylor. Learning representations for pixel-based control: What matters and why? *arXiv preprint arXiv:2111.07775*, 2021.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

Claas Voelcker, Tyler Kastner, Igor Gilitschenski, and Amir-massoud Farahmand. When does self-prediction help? understanding auxiliary tasks in reinforcement learning. *arXiv preprint arXiv:2406.17718*, 2024.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021a.

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, pp. 10674–10681, 2021b.

Hongyu Zang, Xin Li, and Mingzhong Wang. Simsr: Simple distance-based state representations for deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8997–9005, 2022.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.