Self-Consistent Model-based Adaptation for Visual Reinforcement Learning

Xinning Zhou, **Chengyang Ying**, **Yao Feng**, **Hang Su*** and **Jun Zhu***
Department of Computer Science & Technology, Institute for AI, BNRist Center,
Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University
zxn21@mails.tsinghua.edu.cn, dcszj@tsinghua.edu.cn

Abstract

Visual reinforcement learning agents typically face serious performance declines in real-world applications caused by visual distractions. Existing methods rely on fine-tuning the policy's representations with hand-crafted augmentations. In this work, we propose Self-Consistent Model-based Adaptation (SCMA), a novel method that fosters robust adaptation without modifying the policy. By transferring cluttered observations to clean ones with a denoising model, SCMA can mitigate distractions for various policies as a plug-and-play enhancement. To optimize the denoising model in an unsupervised manner, we derive an unsupervised distribution matching objective with a theoretical analysis of its optimality. We further present a practical algorithm to optimize the objective by estimating the distribution of clean observations with a pre-trained world model. Extensive experiments on multiple visual generalization benchmarks and real robot data demonstrate that SCMA effectively boosts performance across various distractions and exhibits better sample efficiency.

1 Introduction

Visual reinforcement learning (VRL) aims to complete complex tasks with high-dimensional observations, which has achieved remarkable results in various domains [Hafner *et al.*, 2019a; Brohan *et al.*, 2023; Li *et al.*, 2024]. Since VRL agents are typically trained on clean observations with minimal distractions, they struggle to handle cluttered observations when deployed in real-world environments with unexpected visual distractions, such as changes in textures or complex backgrounds [Hansen *et al.*, 2020; Fu *et al.*, 2021]. The discrepancy between clean and cluttered observations results in a serious performance gap.

The key to closing the performance gap is to make the policy invariant to distractions. Most existing methods aim to mitigate distractions by learning robust representations. In particular, one line of work is to align the policy's representation between the clean and cluttered observations. Due to the lack of paired data, prevailing methods use hand-crafted functions to create augmentations similar to cluttered observations [Hansen and Wang, 2021; Bertoin *et al.*, 2022]. The effectiveness of such methods is typically limited in settings without prior knowledge of potential distractions. Another line of work addresses the problem through adaptation, which boosts deployment performance by fine-tuning the policy's representation with self-supervised objectives. However, existing adaptation-based methods often lead to narrow empirical increases [Hansen *et al.*, 2020] or are effective only for a specific type of distractions [Yang *et al.*, 2024]. Moreover, the practical application of VRL often requires different policies to ensure robustness against the same types of distractions [Devo *et al.*, 2020]. For instance, domestic robots for different tasks all face the challenge imposed by residential backgrounds with similar distractions. Since policies trained for different tasks have distinct representations, current methods need to fine-tune them separately as the modification made to one policy's representations is not directly applicable to another policy.

To address the above issues, we propose **S**elf-**C**onsistent **M**odel-based **A**daptation (SCMA), a novel method that fosters robust adaptation for various policies as a plug-and-play enhancement. Instead of fine-tuning policies, SCMA utilizes a denoising model to mitigate distractions by transferring cluttered observations to clean ones. Therefore, the denoising model is policy-agnostic and can be seamlessly combined with any policy to boost performance under distractions without modifying its parameters. We further design an unsupervised distribution matching objective to optimize the denoising model in the absence of paired data. Theoretically, we show that the solution set of the unsupervised objective strictly contains the optimal solution in the supervised setting. The proposed objective regularizes the outputs of the denoising model to follow the distribution of observations in clean environments, which we choose to estimate with a pre-trained world model [Hafner *et al.*, 2019b; Hafner *et al.*, 2023].

We practically evaluate the effectiveness of SCMA with the commonly adopted DMControlGB [Hansen *et al.*, 2020; Hansen and Wang, 2021], DMControlView [Yang *et al.*, 2024], and RL-ViGen [Yuan *et al.*, 2024], where the agent needs to complete continuous control tasks in environments with visual distractions. Extensive results show that SCMA

*Corresponding author.

significantly narrows the performance gap caused by various types of distractions, including natural video background, moving camera view, and occlusion. Also, we verify the effectiveness of SCMA with real-world robot data, showing its future potential in real-world deployment. In summary, the main contributions of this paper are:

- We address the challenge of visual distractions by transferring observations and derive an unsupervised distribution matching objective with theoretical analysis.
- We propose self-consistent model-based adaptation (SCMA), a novel method that promotes robust adaptation for different policies in a plug-and-play manner.
- Extensive experiments show that SCMA significantly closes the performance gap caused by various types of distractions. We also demonstrate the effectiveness of SCMA with real-world robot data.

2 Related Work

2.1 Visual Generalization in RL

The ability to generalize across environments with unknown distractions is a long-standing challenge for the practical application of reinforcement learning (RL) agents [Chaplot *et al.*, 2020; Tomar *et al.*, 2021; Liu *et al.*, 2023]. Task-induced methods address the problem by learning structured representations that separate task-relevant features from confounding factors [Fu *et al.*, 2021; Pan *et al.*, 2022; Wang *et al.*, 2022]. Augmentation-based methods regularize the representation between augmented images and clean equivalents [Hansen and Wang, 2021; Ha *et al.*, 2023], but they require prior knowledge of the test-time variations to manually design augmentations. Adaptation-based methods [Hansen *et al.*, 2020; Yang *et al.*, 2024] do not assume the distractions and fine-tune the agent's representation through self-supervised objectives. However, existing adaptation-based methods tend to lead to narrow empirical improvement [Hansen *et al.*, 2020] or are limited to a specific type of visual distractions [Yang *et al.*, 2024]. Several studies aim to tackle this issue with foundation models [Nair *et al.*, 2022; Shah *et al.*, 2023], but they still struggle with computational budget and inference time.

2.2 Unsupervised Domain Transfer

Unsupervised Domain Transfer aims to map data collected from the source domain to a related target domain without explicit supervision signals [Wang *et al.*, 2021]. The topic has been explored in various research areas, such as style transfer [Zhu *et al.*, 2017; Zhao *et al.*, 2022], pose transfer [Li *et al.*, 2023], language translation [Lachaux *et al.*, 2020; Artetxe *et al.*, 2017] and so on. However, one key difference between our setting and theirs is that we can interact with the environments to collect data rather than relying on precollected static datasets. Therefore, we can obtain a certain level of control over the distribution of collected data by selecting specific action sequences, which makes it possible for us to achieve the desired transfer from cluttered observations to clean ones with unsupervised distribution matching [Cao *et al.*, 2018; Baktashmotlagh *et al.*, 2016].

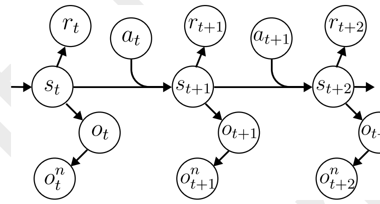

Figure 1: The graphical model of a NPOMDP, where o_t and o_t^n denote the clean and cluttered observation respectively.

3 Methodology

We first present our problem formulation and the supervised objective $\mathcal{L}_O$ in Sec. 3.1. Then we introduce an unsupervised distribution matching surrogate $\mathcal{L}_{\text{KL}}$ and analyze the connection between $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_O$ in Sec. 3.2. Finally, we transform $\mathcal{L}_{\text{KL}}$ into several optimizable adaptation losses in Sec. 3.3, along with practical enhancements in Sec. 3.4.

3.1 Problem Formulation

We formalize visual RL with distractions as a Noisy Partially-Observed Markov Decision Process (NPOMDP) $\mathcal{M}_n = \langle \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, \rho_0, f_n \rangle$. In a NPOMDP, $\mathcal{S}$ is the hidden state space, $\mathcal{O}$ is the discrete observation space, $\mathcal{A}$ denotes the action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ defines the transition probability distribution over the next state, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, γ is the discount factor, and ρ_0 is the initial state distribution. Here $f_n : \mathcal{O} \mapsto \mathcal{O}$ is a noise function that maps a clean observation o_t to its cluttered version $o_t^n = f_n(o_t)$. Following the common settings [Hansen *et al.*, 2020; Bertoin *et al.*, 2022], we assume that f_n is injective so that the distractions do not corrupt the original information. The graphical model of NPOMDP is provided in Fig. 1.

Given the action sequence $a_{1:T}$, the conditional joint distribution describing the environment's latent dynamics is defined as:

$$p(o_{1:T}, o^n_{1:T}, r_{1:T} | a_{1:T}) \coloneqq \int \prod_{t=1}^{T} p(o_t^n|o_t) p(o_t|s_{\le t}, a_{<t}) p(r_t|s_{\le t}, a_{<t}) p(s_t|s_{<t}, a_{<t}) \mathrm{d}s_{1:T}. \tag{1}$$

We denote $p(o_t^n|o_t) = \delta(o_t^n - f_n(o_t))$ as the *noising distribution* of f_n, which is a Dirac distribution with $\delta(\cdot)$ being the Dirac delta function [Dirac, 1981]. Leveraging the Bayes' rule, the posterior distribution $p(o_t|o_t^n)$ can also be derived from Eq. 1, which we denote as the *posterior denoising distribution* of f_n.

The performance of policies pre-trained with clean observations often degenerates when handling cluttered observations [Hansen *et al.*, 2020; Bertoin *et al.*, 2022]. To fill the performance gap, a natural way is to transfer cluttered observations to their corresponding clean ones by estimating the posterior denoising distribution $p(o_t|o_t^n)$. In the supervised setting, we can estimate $p(o_t|o_t^n)$ with a learnable distribution $q(o_t|o_t^n)$ by minimizing the following objective:

$$\begin{aligned} \mathcal{L}_O &\coloneqq -\mathbb{E}_{p(o_{1:T}, o^n_{1:T}|a_{1:T})} \log q(o_{1:T}|o^n_{1:T}) \\ &= -\mathbb{E}_{p(o_{1:T}, o^n_{1:T}|a_{1:T})} \sum_t \log q(o_t|o_t^n). \end{aligned}$$

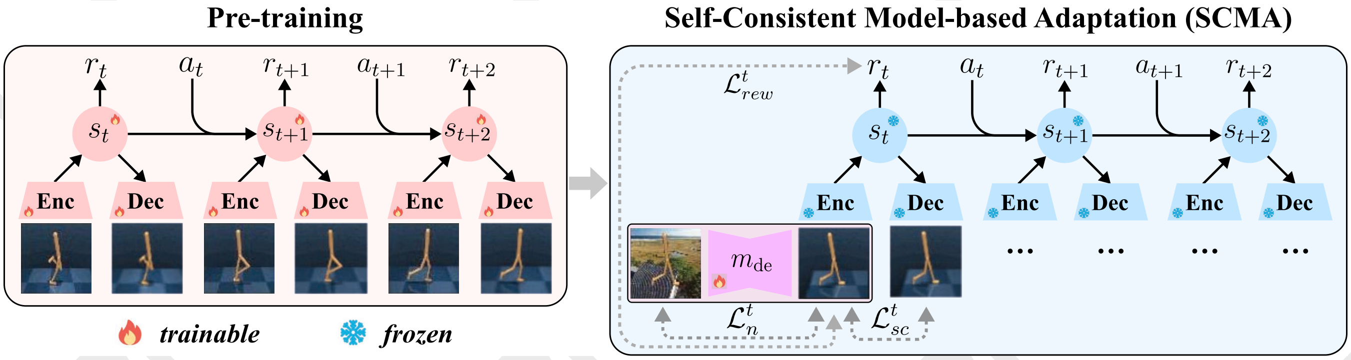

Figure 2: **An overview of Self-Consistent Model-based Adaptation (SCMA)**. SCMA adapts the agent to distracting environments by transferring cluttered observations to clean ones with the denoising model m_{de}. Leveraging a pre-trained world model, m_{de} can be efficiently optimized with self-consistent reconstruction, noisy reconstruction, and reward prediction loss.

We further show that $p(o_t|o_t^n)$ is a Dirac distribution when f_n is injective. Therefore, we adopt a denoising model m_{de} and choose $q(o_t|o_t^n) = \delta(o_t - m_{\text{de}}(o_t^n))$ in practice. More details can be found in Appendix A.1.

3.2 Mitigating Visual Distractions with Unsupervised Distribution Matching

The direct optimization of $\mathcal{L}_{\mathcal{O}}$ requires collecting paired observations (o_t, o_t^n). Since we can only collect observations from clean environments (i.e. $p(o_{1:T}|a_{1:T})$) and distracting environments (i.e. $p(o_{1:T}^n|a_{1:T})$) separately, the absence of paired data imposes severe challenges. Inspired by unsupervised distribution matching [Baktashmotlagh *et al.*, 2016; Cao *et al.*, 2018], we propose to minimize the KL-divergence between the action-conditioned distribution of the clean and transferred observations, which leads to the following unsupervised surrogate $\mathcal{L}_{\text{KL}}$ (see Appendix A.2 for details):

$$\mathcal{L}_{\text{KL}} \coloneqq \mathrm{D}_{\text{KL}}\Big(p(o_{1:T}^n|a_{1:T})q(o_{1:T}|o_{1:T}^n) \Big\| p(o_{1:T}|a_{1:T})q(o_{1:T}^n|o_{1:T})\Big),$$

where $q(o_{1:T}|o_{1:T}^n) = \prod_t q(o_t|o_t^n)$ and $q(o_{1:T}^n|o_{1:T}) = \prod_t q(o_t^n|o_t)$ are learnable denoising and noisy distribution respectively.

To analyze the connection between $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\mathcal{O}}$, we first introduce the concept of *homogeneous noise functions*, which are theoretically indistinguishable in the unsupervised setting and defined below (Details are deferred to Appendix A.2):

Definition 1. *For noise functions f_{n_i}, we denote $o_t^{n_i} = f_{n_i}(o_t)$ as its cluttered observation. Given the distribution of clean observation $p(o_{1:T}|a_{1:T})$, we call the noise functions f_{n_1} and f_{n_2} to be homogeneous under $p(o_{1:T}|a_{1:T})$ if their cluttered observations have the same distribution, i.e.:*

$$f_{n_1} \equiv_p f_{n_2} \Leftrightarrow p(o_{1:T}^{n_1}|a_{1:T}) = p(o_{1:T}^{n_2}|a_{1:T}),$$

$$\textit{where } p(o_{1:T}^{n_i}|a_{1:T}) = \sum_{o_{1:T}} p(o_{1:T}|a_{1:T})p(o_{1:T}^{n_i}|o_{1:T}).$$

We define $\mathcal{H}_{f_n}^p = \{f_{n_i}|f_{n_i} \equiv_p f_n\}$, which includes all homogeneous noise functions of f_n under $p(o_{1:T}|a_{1:T})$.

Then we show that the solution set of $\mathcal{L}_{\text{KL}}$ equals the set of posterior denoising distributions of noise functions in $\mathcal{H}_{f_n}^p$. Since f_n is clearly in $\mathcal{H}_{f_n}^p$, the solution set of $\mathcal{L}_{\text{KL}}$ contains $p(o_t|o_t^n)$, which is the optimal solution to $\mathcal{L}_{\mathcal{O}}$.

Theorem 1 (Proof in Appendix A.3). *Given $p(o_{1:T}|a_{1:T})$ and $p(o_{1:T}^n|a_{1:T})$, let $\mathcal{Q}$ denote the solution set of $\mathcal{L}_{\text{KL}}$:*

$$\mathcal{Q} \coloneqq \operatorname*{arg\,min}_{q(o_t|o_t^n)} \min_{q(o_t^n|o_t)} \mathcal{L}_{\text{KL}}.$$

It follows that $\mathcal{Q}$ equals the set of posterior denoising distributions of noise functions in $\mathcal{H}_{f_n}^p$:

$$\mathcal{Q} = \left\{p(o_t|o_t^{n_i})|f_{n_i} \in \mathcal{H}_{f_n}^p\right\}. \tag{2}$$

Generally speaking, since homogeneous noise functions are theoretically indistinguishable in the unsupervised setting, we can only assure that m_{de} learns to transfer cluttered observations back to clean ones according to a noise function in $\mathcal{H}_{f_n}^p$. In Appendix A.3, we further reveal the relationship between the number of homogeneous noise functions and properties of $p(o_{1:T}|a_{1:T})$. We also discuss possible ways to reduce the number of homogeneous noise functions in Sec. 3.4 so that $\mathcal{Q}$ only contains $p(o_t|o_t^n)$.

To simplify the computation, we show in Appendix A.2 that $\mathcal{L}_{\text{KL}}$ leads to the following objective, where C is a constant:

$$\begin{aligned}\mathcal{L}_{\text{KL}} &= \mathbb{E}_{p(o_{1:T}^n|a_{1:T})}\Big[\mathrm{D}_{\text{KL}}\big(q(o_{1:T}|o_{1:T}^n)\|p(o_{1:T}|a_{1:T})\big) \\ &\qquad - \mathbb{E}_{q(o_{1:T}|o_{1:T}^n)}[\log q(o_{1:T}^n|o_{1:T})]\Big] + C \\ &= \mathbb{E}_{p(o_{1:T}^n|a_{1:T})}\mathbb{E}_{q(o_{1:T}|o_{1:T}^n)}\Big[-\log p(o_{1:T}|a_{1:T}) \\ &\qquad - \log q(o_{1:T}^n|o_{1:T})\Big] + C.\end{aligned} \tag{3}$$

Intuitively, the first term regularizes the transferred observations to follow the clean environments' latent dynamics $p(o_{1:T}|a_{1:T})$. The second term ensures that the transferred observations remain relevant to the cluttered observations and thus preserve necessary information.

3.3 Adaptation with Pre-trained World Models

Based on the above analyses, we now present the Self-Consistent Model-based Adaptation (SCMA) method, a practical adaptation algorithm that mitigates distractions by optimizing the denoising model with Eq. 3.

Specifically, Eq. 3 involves calculating the action-conditioned distribution $p(o_{1:T}|a_{1:T})$, which we estimate with a pre-trained world model [Hafner *et al.*, 2019b; Hafner *et al.*, 2023]. Given a clean trajectory $\tau = \{o_1, a_1, \cdots, o_T, a_T\}$, the world model estimates $\log p(o_{1:T}|a_{1:T})$ with $\log p_{\text{wm}}(o_{1:T}|a_{1:T})$ by maximizing the following evidence lower bound (ELBO):

$$\begin{aligned}\log p_{\text{wm}}(o_{1:T}|a_{1:T}) &= \log \int p_{\text{wm}}(o_{1:T}, s_{1:T}|a_{1:T}) \mathrm{d}s_{1:T} \\ \geq \sum_{t=1}^{T} \mathbb{E}_{q_{\text{wm}}(s_{1:T}|a_{1:T},o_{1:T})} &[\underbrace{\log p_{\text{wm}}(o_t|s_{\leq t}, a_{<t})}_{\mathcal{J}_o^t} \\ &- \underbrace{\mathrm{D_{KL}}(q_{\text{wm}}(s_t|s_{<t}, a_{<t}, o_t) \| p_{\text{wm}}(s_t|s_{<t}, a_{<t}))}_{\mathcal{J}_{kl}^t}].\end{aligned} \tag{4}$$

In the above objective, the KL-divergence objective $\mathcal{J}_{kl}^t$ enables the model's generation ability by minimizing the distance between the prior and posterior distribution. The reconstruction objective $\mathcal{J}_o^t$ enforces the model to capture the visual essence of the task by predicting the subsequent observations, which facilitates the later adaptation.

Self-consistent Model-based Adaptation. Before adaptation, we first pre-train the policy and world model in clean environments. Then we deploy the pre-trained policy and our denoising model into the distracting environment to collect trajectory $\{o_1^n, a_1, \cdots, o_T^n, a_T\}$, where actions are sampled as $a_t \sim \pi(\cdot|m_{\text{de}}(o_t^n))$. By estimating $p(o_{1:T}|a_{1:T})$ with the pre-trained world model, optimizing Eq. 3 leads to the following self-consistent reconstruction loss $\mathcal{L}_{sc}^t$ and noisy reconstruction loss $\mathcal{L}_n^t$. It should be noted that the policy and world model are both frozen during adaptation. We choose to drop a similar KL-loss term as in Eq. 4 because we empirically find it to have a negative impact on adaptation by harming the reconstruction results, consistent with previous works [Higgins *et al.*, 2017; Chen *et al.*, 2018]. The detailed derivation is provided in Appendix A.4.

$$\mathcal{L}_{sc}^t = -\mathbb{E}_{q(o_{1:T}|o_{1:T}^n)} \mathbb{E}_{q_{\text{wm}}(s_{1:T}|a_{1:T},o_{1:T})} [\log p_{\text{wm}}(o_t|s_{\leq t}, a_{<t})],$$
$$\mathcal{L}_n^t = -\mathbb{E}_{q(o_{1:T}|o_{1:T}^n)} [\log q(o_t^n|o_t)].$$

$\mathcal{L}_{sc}^t$ encourages the denoising model to transfer cluttered observations to clean ones so that the transferred observations will conform to the prediction of the world model. $\mathcal{L}_n^t$ prevents the denoising model from ignoring the cluttered observations and thus outputting clean yet irrelevant observations. In practice, we implement $q(o_t^n|o_t) = \delta(o_t^n - m_{\text{n}}(o_t))$ with a noisy model m_{n}, and $q(o_{1:T}|o_{1:T}^n) = \prod_t q(o_t|o_t^n) = \prod_t \delta(o_t - m_{\text{de}}(o_t^n))$ with a denoising model m_{de}.

3.4 Boosting Adaptation by Reducing Homogeneous Noise Functions

As discussed in Theorem 1, the solution set of $\mathcal{L}_{\text{KL}}$ equals the set of posterior denoising distributions of noise functions in $\mathcal{H}_{f_n}^p$. To promote the adaptation, we propose two practical techniques to help the denoising distribution $q(o_t|o_t^n)$ converge to the target posterior denoising distribution $p(o_t|o_t^n)$ by reducing the number of homogeneous noise functions.

Leverage Rewards. If reward signals are available in distracting environments, they can naturally boost adaptation by reducing the number of homogeneous noise functions. Loosely speaking, noise functions with the same $p(o_{1:T}^n|a_{1:T})$ but different $p(o_{1:T}^n, r_{1:T}|a_{1:T})$ are no longer homogeneous if rewards are available. A detailed explanation is provided in Appendix A.3. The derivation in Sec. 3.2 can be simply extended to include rewards by redefining $\mathcal{L}_{\text{KL}}$ as below (details in Appendix A.4):

$$\begin{aligned}\mathcal{L}_{\text{KL}} \coloneqq \mathrm{D_{KL}}\Big(&p(o_{1:T}^n, r_{1:T}|a_{1:T}) q(o_{1:T}|o_{1:T}^n) \\ &\| p(o_{1:T}, r_{1:T}|a_{1:T}) q(o_{1:T}^n|o_{1:T})\Big),\end{aligned}$$

which leads to the reward prediction loss:

$$\mathcal{L}_{rew}^t = -\mathbb{E}_{q(o_{1:T}|o_{1:T}^n)} \mathbb{E}_{q_{\text{wm}}(s_{1:T}|a_{1:T},o_{1:T})} [\log p_{\text{wm}}(r_t|s_{\leq t}, a_{<t})].$$

$\mathcal{L}_{rew}^t$ encourages the transferred observations to contain sufficient information of rewards and ignore reward-irrelevant distractions. The final adaptation loss of SCMA is:

$$\mathcal{L}_{\text{SCMA}}^t = \mathcal{L}_{sc}^t + \mathcal{L}_n^t + \mathcal{L}_{rew}^t. \tag{5}$$

Limit the Hypothesis Set of the Denoising Model. For specific types of distractions, we can further encode some inductive bias in the denoising model architecture. Therefore, we can prevent $q(o_t|o_t^n)$ from converging to the posterior denoising distributions of certain homogeneous noise functions by limiting the hypothesis set. For example, we can implement the denoising model as a mask model $m_{\text{mask}} : \mathbb{R}^{h \times w \times c} \mapsto [0, 1]^{h \times w \times c}$ to handle background distractions. However, to verify the generality of SCMA, we refrain from assuming the type of distractions and implement the denoising model as a generic image-to-image network by default. Detailed implementations are provided in Appendix C.2.

In summary, we propose an adaptation framework with a two-stage pipeline: 1) pre-training the policy and world model in clean environments to master skills and capture the environments' latent dynamics $p(o_{1:T}|a_{1:T})$. 2) adapting the policy to visually distracting environments by optimizing $q(o_t|o_t^n)$ with Eq. 5 to transfer cluttered trajectories to clean ones. The pipeline is illustrated with Fig. 2, with further training details provided in Appendix C.

4 Experiment

In this section, we evaluate the capability of SCMA by addressing the following questions:

- Can SCMA fill the performance gap caused by various types of distractions?
- Can SCMA generalize across various tasks or policies from different algorithms?
- How does each loss component contribute to the results? Can SCMA still handle distractions without rewards?
- Can SCMA converge faster compared to other adaptation-based methods or directly training from scratch in visually distracting environments?

video_hard	SCMA	SCMA (w/o r)	MoVie	PAD	SVEA	Dr. G	SGQN	TIA	TPC	DreamerPro
ball_in_cup-catch	**809$\pm$114**	215$\pm$60	41$\pm$20	130$\pm$47	498$\pm$147	635$\pm$26	782$\pm$57	329$\pm$466	220$\pm$207	378$\pm$231
cartpole-swingup	**773$\pm$51**	145$\pm$40	83$\pm$2	123$\pm$21	401$\pm$38	545$\pm$23	544$\pm$43	98$\pm$22	219$\pm$19	365$\pm$48
finger-spin	**948$\pm$5**	769$\pm$182	2$\pm$0	96$\pm$11	307$\pm$24	-	822$\pm$24	146$\pm$93	315$\pm$40	427$\pm$299
walker-stand	**953$\pm$4**	328$\pm$30	127$\pm$23	336$\pm$22	747$\pm$43	-	851$\pm$24	117$\pm$9	840$\pm$98	941$\pm$14
walker-walk	722$\pm$89	129$\pm$19	39$\pm$13	108$\pm$33	385$\pm$63	**782$\pm$37**	739$\pm$21	84$\pm$55	402$\pm$57	617$\pm$159

(a) video_hard

moving_view	SCMA	MoVie	PAD	SGQN
ball_in_cup-catch	745$\pm$121	**951$\pm$10**	750$\pm$32	857$\pm$64
cartpole-swingup	708$\pm$76	167$\pm$25	561$\pm$86	**788$\pm$65**
finger-spin	**952$\pm$10**	896$\pm$21	603$\pm$28	702$\pm$56
walker-stand	**977$\pm$16**	712$\pm$11	955$\pm$15	961$\pm$2
walker-walk	**922$\pm$55**	810$\pm$7	645$\pm$21	769$\pm$36

(b) moving_view

color_hard	SCMA	MoVie	PAD	SGQN	SVEA
ball_in_cup-catch	817$\pm$64	67$\pm$41	563$\pm$50	881$\pm$61	**961$\pm$7**
cartpole-swingup	809$\pm$15	102$\pm$14	630$\pm$63	773$\pm$80	**837$\pm$23**
finger-spin	965$\pm$2	652$\pm$10	803$\pm$72	847$\pm$80	**977$\pm$5**
walker-stand	**984$\pm$11**	121$\pm$14	797$\pm$46	867$\pm$81	942$\pm$26
walker-walk	**954$\pm$7**	38$\pm$3	468$\pm$74	828$\pm$84	760$\pm$145

(c) color_hard

occlusion	SCMA	MoVie	PAD	SGQN
ball_in_cup-catch	**899$\pm$41**	33$\pm$18	145$\pm$6	642$\pm$74
cartpole-swingup	**779$\pm$10**	120$\pm$32	142$\pm$9	127$\pm$18
finger-spin	**920$\pm$1**	1$\pm$0	15$\pm$9	117$\pm$22
walker-stand	**976$\pm$17**	124$\pm$21	305$\pm$16	376$\pm$87
walker-walk	**902$\pm$51**	52$\pm$15	94$\pm$24	118$\pm$34

(d) occlusion

RL-ViGen	SCMA	SGQN	SRM	SVEA	CURL
Door (easy)	**416$\pm$26**	391$\pm$95	337$\pm$110	268$\pm$136	6$\pm$5
Door (extreme)	**380$\pm$30**	160$\pm$122	31$\pm$18	62$\pm$56	2$\pm$1
Lift (easy)	19$\pm$5	31$\pm$17	**69$\pm$32**	43$\pm$18	0$\pm$0
Lift (extreme)	**15$\pm$9**	7$\pm$7	0$\pm$0	8$\pm$5	0$\pm$0
TwoArm (easy)	340$\pm$27	349$\pm$23	**419$\pm$45**	414$\pm$58	150$\pm$20
TwoArm (extreme)	227$\pm$24	**257$\pm$31**	161$\pm$27	155$\pm$18	147$\pm$15

(e) Table-top Manipulation tasks in RL-ViGen.

Table 1: Performance (mean $\pm$ std) in visually distracting environments. We report the performance of SCMA and baseline methods in DMControl and RL-ViGen across various distracting settings. The best algorithm is **bold** for every task.

4.1 Experiment Setup

Environments. To measure the effectiveness of SCMA, we follow the settings from the commonly adopted DMControlGB [Hansen and Wang, 2021; Hansen *et al.*, 2021; Bertoin *et al.*, 2022], DMControlView [Yang *et al.*, 2024], and RL-ViGen [Yuan *et al.*, 2024]. The agent is asked to perform continuous control tasks in visually distracting environments, including video distracting background (video_hard), moving camera views (moving_view), and randomized colors (color_hard). We also evaluate the agent's performance in a more challenging occlusion setting by randomly masking $1/4$ of each observation. We provide a visualization of every distracting environment in Fig. 6 in the Appendix. Unless otherwise stated, the result of each task is evaluated over 3 seeds and we report the average performance of the policy in the last episode.

Baselines. We compare SCMA to the state-of-the-art adaptation-based baselines: PAD [Hansen *et al.*, 2020], MoVie [Yang *et al.*, 2024]. We also include comparison with other kinds of methods, including augmentation-based methods: SVEA [Hansen *et al.*, 2021], SGQN [Bertoin *et al.*, 2022], Dr. G [Ha *et al.*, 2023]; and task-induced methods: TIA [Fu *et al.*, 2021], TPC [Nguyen *et al.*, 2021], DreamerPro [Deng *et al.*, 2022]. Following the official design [Hansen and Wang, 2021], the augmentation-based methods use random overlay with images from Place365 [Zhou *et al.*, 2017]. Task-induced methods directly learn the structured representations in distracting environments. Adaptation-based methods will first be pre-trained in the clean environments for 1M timesteps and then adapt to the distracting environments for 0.1M timesteps (0.4M for video_hard and 0.5M for RL-ViGen). By default, SCMA adapts a pre-trained Dreamer policy [Hafner *et al.*, 2019a] to distracting environments. More details can be found in Appendix C.1.

4.2 Adaptability to Visual Distractions

We first evaluate the adaptation ability of SCMA by measuring its performance in the challenging visual generalization benchmarks. Before adapting to the visually distracting environments, we first pre-train the policy and world model in the clean training environment (see Fig. 6a in the Appendix). Then we adapt the agent to visually distracting environments leveraging the pre-trained world model. The experiment results in Table 1 show that SCMA significantly reduces the performance gap caused by distractions and achieves appealing performance compared to augmentation-based methods. While remaining competitive in the color_hard setting, SCMA outperforms the best baseline method in most tasks in other 3 settings. Moreover, SCMA obtains the best performance in all tasks in the occlusion setting, which is a common scenario for real-world robot controls. To verify the idea of boosting adaptation by reducing the hypothesis set (Sec. 3.4), we implement the denoising model with specific architectures and conduct experiments in the table-top

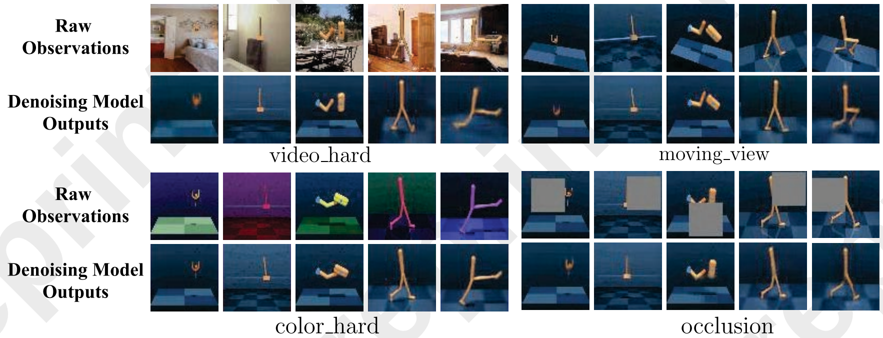

Figure 3: Visualization of the raw observations and the denoising model's outputs in various distracting environments.

occlusion	SGQN	SGQN+SCMA
ball_in_cup-catch	642±74	775±151
cartpole-swingup	127±18	337±51
finger-spin	117±22	133±19
walker-stand	376±87	884±63
walker-walk	118±34	465±101
Averaged	276.0	518.8 (88.0%↑)

Table 2: Performance (mean ± std) in occlusion environment. The results show that the denoising model can boost SGQN's performance in a plug-and-play manner.

manipulation tasks with distracting settings from RL-ViGen. Further details are included in Appendix C.2. Following previous work [Yuan *et al.*, 2024], we report the scores under the eval_easy and eval_extreme settings in Table 1e. The results show that SCMA achieves the best performance in half of the scenarios and remains comparable to other methods in the remaining ones. We believe one way to further improve the performance is to incorporate stronger world models [Ding *et al.*, 2024], which we leave to future works.

We visualize how m_{de} transfers cluttered observations to clean ones in Fig. 3, where it effectively mitigates various types of distractions and restores the task-relevant objects correctly. The qualitative results also indicate that our method can effectively handle distractions not only for large embodiments like walker-walk, but also for challenging small embodiments such as ball_in_cup-catch and cartpole-swingup, which task-induced methods often fail to manage.

4.3 Versatility of the Denoising Model

We conduct experiments to measure the versatility of the denoising model from two aspects: 1) can the denoising model generalize across tasks with the same robot? 2) is the denoising model applicable to policies from different algorithms? To answer the above questions, we first cross-evaluate the capability of the denoising model between walker-walk and walker-stand in the video_hard environment. Specifically, we take the denoising model adapted to one task and directly evaluate its performance in another task. The results in Table 4 in the Appendix indicate that the achieved denoising model is not restricted to a specific task and exhibits appealing zero-shot generalization capability. To verify that the denoising model is agnostic to policies, we first optimize the denoising model with trajectories collected by a Dreamer policy. Then we combine the obtained denoising model with an SGQN policy in a plug-and-play manner and measure the performance in the occlusion setting. While SGQN reaches appealing results in other settings, it performs poorly under occlusions. However, Table 2 demonstrates that incorporating the denoising model can improve the performance of SGQN by 88%. Therefore, SCMA can serve as a convenient component to promote performance under certain distractions without modifying the policy. However, there is a disparity between the performance of SGQN policy with SCMA and Dreamer policy with SCMA, which we attribute to the policy encoder. As the Dreamer policy's encoder leverages the long-term representation extracted by the world model, it is less susceptible to small mistakes made by the denoising model.

4.4 Adaptation Without Rewards

While SCMA utilizes both visual and reward signals for the best adaptation results, the ability to adapt without rewards is also important. To address this issue, we conduct experiments in the video_hard environments to investigate how different loss components affect the final adaptation results.

To better understand the impact of different losses, we separately removed the 3 loss components from SCMA during adaptation, namely self-consistent reconstruction loss $\mathcal{L}_{sc}^t$, reward prediction loss $\mathcal{L}_{rew}^t$, and noisy reconstruction loss $\mathcal{L}_n^t$. From the ablation results in Fig. 15 in the Appendix, we can see that removing the self-consistent reconstruction loss leads to the most significant decrease, indicating that the proposed $\mathcal{L}_{sc}^t$ plays a crucial role in adaptation. Another finding is that the reward loss can promote better adaptation by encouraging the denoising model to focus on some miniature yet critical features, such as the ball in ball_in_cup-catch and the pole in cartpole-balance. While $\mathcal{L}_{rew}^t$ contributes considerably to the final adaptation results, Fig. 8 in the Appendix demonstrates that SCMA without rewards still achieves the highest average performance among all adaptation-based methods. The noisy reconstruction loss mainly aims to preserve the connection between the cluttered and transferred observations. Intuitively, removing $\mathcal{L}_n^t$ from Eq. 3 will cause a *mode-seeking* problem [Cheng, 1995], where the denoising model will prefer the mode of $\log p(o_{1:T}|a_{1:T})$ and thus transfer cluttered observations to

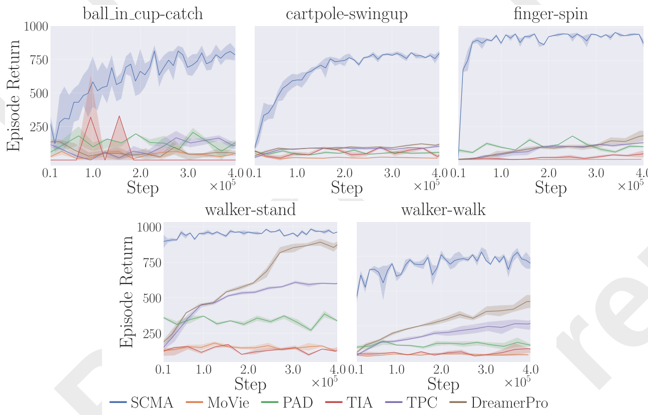

Figure 4: Performance curves of different algorithms in the video_hard environment, where SCMA exhibits better final performance and sample efficiency.

clean yet irrelevant observations.

4.5 Sample Efficiency in Visually Distracting Environments

Accomplishing tasks with as few cluttered observations as possible is practically important to deploy the agent in distracting environments. Compared with other adaptation-based methods or training from scratch with task-induced methods [Fu *et al.*, 2021; Deng *et al.*, 2022], the performance curves in Fig. 4 show that SCMA can achieve higher performance with much fewer downstream cluttered samples. Although we adapt the policy in video_hard with 0.4M steps, SCMA can achieve competitive performance with much fewer steps. We provide the wall clock time and adaptation steps for SCMA to reach 90% of the final performance in Table 5 in the Appendix to show that SCMA obtains compelling results with only 10% of total adaptation timesteps for most tasks.

4.6 Real-world Robot Data

With the rapid development of generative models, their potential to enhance real-world robotic controls has attracted significant attention. Recent works leverage video models to create future observations based on current environment observation and extract executable action sequences with inverse dynamics models (IDM) [Du *et al.*, 2023; Ko *et al.*, 2023]. However, the generated observations might still contain distractions if the input environment observation is cluttered, which imposes challenges to the IDM in making accurate action predictions. We show that SCMA can help IDM better predict the actions when handling cluttered observations. More details are included in Appendix C.3.

We manually collect real-world robot data with a Mobile ALOHA robot by performing an apple-grasping task with teleoperation. The IDM is trained with data collected in the normal setting and evaluated on data collected in 3 distracting settings: 1) fruit_bg: various fruits are placed in the background. 2) color_bg: the scene is disrupted by a blue light. 3) varying_light: the lighting is intentionally changed. We provide the quantitative results in Table 6 in the Appendix and

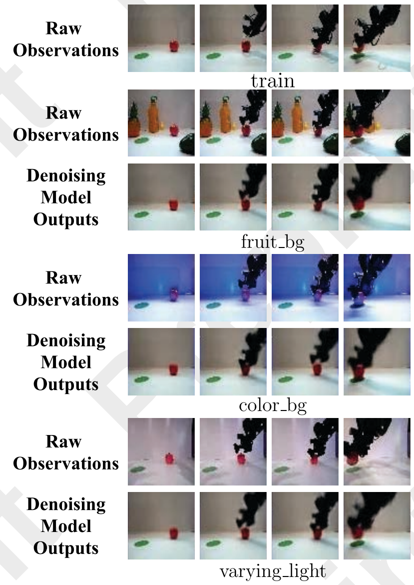

Figure 5: Visualization of the raw observations and denoising model's outputs on real-world robot data.

visualization in Fig. 5. The results show that SCMA effectively mitigates real-world distractions and thus has important implications for the practical deployment of robots.

5 Conclusion and Discussion

The ability to generalize across environments with various distractions is a long-standing goal in visual RL. In this work, we formalize the challenge as an unsupervised transferring problem and propose a novel method called self-consistent model-based adaptation (SCMA). SCMA adopts a policy-agnostic denoising model to mitigate distractions by transferring cluttered observations into clean ones. To optimize the denoising model in the absence of paired data, we propose an unsupervised distribution matching objective that regularizes the outputs of the denoising model to follow the distribution of clean observations, which can be estimated with a pre-trained world model. Experiments in challenging visual generalization benchmarks show that SCMA effectively reduces the performance gap caused by distractions and can boost the performance of various policies in a plug-and-play manner. Moreover, we validate the effectiveness of SCMA with real-world robot data, where SCMA effectively mitigates distractions and promotes better action predictions.

SCMA proposes a general model-based objective for adaptation under distractions, and we wish to further promote this direction by highlighting some limitations and future improvements. SCMA pre-trains world models to estimate the action-conditioned distribution of clean observations. Including stronger world models like diffusion models [Wang *et al.*, 2023] may be a promising way to further promote the performance with complex robots or real-world tasks. Another potential improvement is to explore other types of signals that are invariant between clean and distracting environments, e.g., 3D structures of robots [Driess *et al.*, 2022] or natural language descriptions of tasks [Sumers *et al.*, 2023].

Acknowledgments

This work was supported by NSFC Projects (Nos. 92248303, 92370124, 62350080, 62276149, 62061136001, 92270001), BNRist (BNR2022RC01006), Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. J. Zhu was also supported by the XPlorer Prize. Xinning Zhou and Chengyang Ying contributed equally to this work.

References

[Artetxe *et al.*, 2017] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

[Baktashmotlagh *et al.*, 2016] Mahsa Baktashmotlagh, Mehrtash Har, Mathieu Salzmann, et al. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17(108):1–30, 2016.

[Bertoin *et al.*, 2022] David Bertoin, Adil Zouitine, Mehdi Zouitine, and Emmanuel Rachelson. Look where you look! saliency-guided q-networks for generalization in visual reinforcement learning. In *Neural Information Processing Systems*, 2022.

[Brohan *et al.*, 2023] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR, 2023.

[Cao *et al.*, 2018] Yue Cao, Mingsheng Long, and Jianmin Wang. Unsupervised domain adaptation with distribution matching machines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Caron *et al.*, 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

[Chaplot *et al.*, 2020] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33, 2020.

[Chen *et al.*, 2018] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[Cheng, 1995] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[Deng *et al.*, 2022] Fei Deng, Ingook Jang, and Sungjin Ahn. Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 4956–4975. PMLR, 2022.

[Devo *et al.*, 2020] Alessandro Devo, Giacomo Mezzetti, Gabriele Costante, Mario L Fravolini, and Paolo Valigi. Towards generalization in target-driven visual navigation by using deep reinforcement learning. *IEEE Transactions on Robotics*, 36(5):1546–1561, 2020.

[Ding *et al.*, 2024] Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model: Future modeling beyond step-by-step rollout for offline reinforcement learning, 2024.

[Dirac, 1981] Paul Adrien Maurice Dirac. *The principles of quantum mechanics*. Number 27. Oxford university press, 1981.

[Driess *et al.*, 2022] Danny Driess, Ingmar Schubert, Pete Florence, Yunzhu Li, and Marc Toussaint. Reinforcement learning with neural radiance fields. *Advances in Neural Information Processing Systems*, 35:16931–16945, 2022.

[Du *et al.*, 2023] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *ArXiv*, abs/2302.00111, 2023.

[Fu *et al.*, 2021] Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *International Conference on Machine Learning*, pages 3480–3491. PMLR, 2021.

[Ha *et al.*, 2023] Jeongsoo Ha, Kyungsoo Kim, and Yusung Kim. Dream to generalize: zero-shot model-based reinforcement learning for unseen visual distractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7802–7810, 2023.

[Hafner *et al.*, 2019a] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[Hafner *et al.*, 2019b] 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.

[Hafner *et al.*, 2023] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[Hansen and Wang, 2021] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13611–13617. IEEE, 2021.

[Hansen *et al.*, 2020] Nicklas Hansen, Yu Sun, P. Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. *ArXiv*, abs/2007.04309, 2020.

[Hansen *et al.*, 2021] Nicklas Hansen, Hao Su, and Xiaolong Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. 2021.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[Ko *et al.*, 2023] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Josh Tenenbaum. Learning to act from actionless videos through dense correspondences. *ArXiv*, abs/2310.08576, 2023.

[Lachaux *et al.*, 2020] Marie-Anne Lachaux, Baptiste Roziere, Lowik Chanussot, and Guillaume Lample. Unsupervised translation of programming languages. *arXiv preprint arXiv:2006.03511*, 2020.

[Li *et al.*, 2023] Tianyu Li, Hyunyoung Jung, Matthew Gombolay, Yong Kwon Cho, and Sehoon Ha. Crossloco: Human motion driven control of legged robots via guided unsupervised reinforcement learning. *ArXiv*, abs/2309.17046, 2023.

[Li *et al.*, 2024] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). 2024.

[Liu *et al.*, 2023] Xin Liu, Yaran Chen, Haoran Li, Boyu Li, and Dongbin Zhao. Cross-domain random pre-training with prototypes for reinforcement learning. *arXiv preprint arXiv:2302.05614*, 2023.

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

[Nguyen *et al.*, 2021] 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.

[Pan *et al.*, 2022] Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Isolating and leveraging controllable and noncontrollable visual dynamics in world models. *arXiv preprint arXiv:2205.13817*, 2022.

[Shah *et al.*, 2023] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[Sumers *et al.*, 2023] Theodore R. Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling internet-scale vision-language models into embodied agents. *ArXiv*, abs/2301.12507, 2023.

[Tassa *et al.*, 2018] 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.

[Tomar *et al.*, 2021] 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.

[Wang *et al.*, 2021] Feng Wang, Lianmeng Jiao, and Quan Pan. A survey on unsupervised transfer clustering. In *2021 40th Chinese Control Conference (CCC)*, pages 7361–7365. IEEE, 2021.

[Wang *et al.*, 2022] Tongzhou Wang, Simon 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*, pages 22591–22612. PMLR, 2022.

[Wang *et al.*, 2023] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[Yang *et al.*, 2024] Sizhe Yang, Yanjie Ze, and Huazhe Xu. Movie: Visual model-based policy adaptation for view generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yuan *et al.*, 2022] Zhecheng Yuan, Zhengrong Xue, Bo Yuan, Xueqian Wang, Yi Wu, Yang Gao, and Huazhe Xu. Pre-trained image encoder for generalizable visual reinforcement learning. *Advances in Neural Information Processing Systems*, 35:13022–13037, 2022.

[Yuan *et al.*, 2024] Zhecheng Yuan, Sizhe Yang, Pu Hua, Can Chang, Kaizhe Hu, and Huazhe Xu. Rl-vigen: A reinforcement learning benchmark for visual generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhao *et al.*, 2022] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *ArXiv*, abs/2207.06635, 2022.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[Zhu *et al.*, 2020] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.