Domain Adaptation with Adaptive Imagination for Visual Reinforcement Learning under Limited Target Data

Hyunwoo Park[1] and Sang-Hyun Lee[2]

[1] STRADVISION, Seoul, Republic of Korea
hyunwoo.park@stradvision.com
[2] Department of Mobility Engineering, Ajou University, Suwon, Republic of Korea
sanghyunlee@ajou.ac.kr

Abstract. Sim-to-real transfer remains a major obstacle for reinforcement learning (RL), especially for vision-based control where image observations exacerbate the state-distribution shift between simulation and the real world. Domain adaptation (DA) is a promising remedy for this challenge. Prior sim-to-real DA works have demonstrated encouraging results, yet these approaches typically assume substantially more target data, which is not available in practice. Indeed, their performance degrades significantly when the target data budget is reduced. To address this challenge, we propose AIDA (**A**daptive **I**magination for **D**omain **A**daptation), a domain adaptation framework for visual reinforcement learning that addresses sim-to-real transfer under scarce target data without requiring additional interaction with the target environment. Our key idea is adaptive imagination: generating reliable and semantic imagination rollouts to augment limited target data. Specifically, AIDA employs a distribution-shift-aware discriminator that truncates rollouts when imagined transitions drift into low-confidence regions, so that only reliable transitions contribute to the augmentation. On these reliable transitions, AIDA introduces a self-consistency loss that cycles through state $\rightarrow$ image observation $\rightarrow$ state, penalizing discrepancies between the original and reconstructed states. This provides additional adaptation signals beyond the scarce target data. Our experiments demonstrate that adaptive imagination effectively truncates unreliable rollouts. By enforcing a self-consistency loss on the resulting reliable transitions, AIDA learns semantically meaningful state representations and outperforms baselines across five MuJoCo tasks and two Gymnasium-Robotics tasks.

1 Introduction

Deep reinforcement learning (RL) has achieved remarkable success in simulation, enabling agents to master complex control tasks from raw sensory inputs [1,12]. However, deploying these learned policies in the real world remains a fundamental challenge: discrepancies between simulation and reality—known as the sim-to-real gap—often cause dramatic performance degradation [24]. This gap can

be broadly attributed to two factors: dynamics mismatch and state-distribution shift. In particular, for vision-based RL policies, state-distribution shift often plays a larger role. This is because pixel observations are high-dimensional and highly sensitive to changes in appearance and sensing conditions, which exacerbates the sim-to-real gap. In this work, we focus on mitigating this state-distribution shift for vision-based RL policies.

To cope with such distribution shifts, three families of methods are commonly used: domain randomization (DR), domain generalization (DG), and domain adaptation (DA). DR [9,10,15,18] augments the simulator to train a single robust policy, hoping that the resulting variability covers factors of variation in the target environment. However, when the test domain is truly unknown or the shift is large, such coverage cannot be guaranteed. DG [3,7,21] aims to learn policies that transfer to previously unseen test environments without target-domain supervision, typically by encouraging domain-agnostic representations and invariances (e.g., via attention/invariance objectives or augmentation-consistent regularization) rather than explicitly matching to a specific target domain. However, because the test domain is unknown by design, there is no guarantee that the training sources or augmentations capture the true factors of variation encountered at deployment; consequently, when the domain gap is large or includes unmodeled shifts, DG performance may degrade substantially—revealing a fundamental limitation of methods that must anticipate the target distribution without ever observing it [21].

DA [2, 6, 16, 20] is a promising alternative that directly leverages target-domain data to adapt to the specific target environment. This requires access to target data, but in return enables specialization to the particular deployment setting, even when the underlying causes of the shift are unknown at training time. Prior DA methods have indeed demonstrated superior performance over DR and DG [6, 20]. Nevertheless, they typically assume access to sufficient target data, and their performance degrades significantly when this budget is reduced [2, 13]. Moreover, most prior DA methods adopt an image-to-image setting [6, 16, 20], which requires training RL agents in image-based simulation—a process that is usually harder [11, 23] and can slow down RL by up to 20× [22]. These two practical constraints remain underexplored despite their practical relevance.

To address these challenges, we propose **AIDA** (**A**daptive **I**magination for **D**omain **A**daptation), a DA framework for visual RL that operates under scarce target data without additional target interaction. Our key idea is *adaptive imagination*: augmenting limited target data with reliable imagination rollouts obtained via discriminator-gated truncation, and optimizing a self-consistency objective on these rollouts to learn semantically meaningful representations. Specifically, AIDA adopts a cross-modality setting (sim-state → real-image) that avoids the overhead of image-based RL training. A learned dynamics model produces policy-conditioned rollouts starting from target-inferred states, augmenting the limited target data with synthetic transitions. However, longer rollouts accumulate model error and drift from the target manifold. Therefore, a three-way discriminator assesses each imagined transition and truncates the rollout when

confidence drops, so that only reliable, target-aligned transitions contribute to training. This adaptive truncation is critical, as imagination reliability varies across both tasks and individual states, making any fixed imagination horizon suboptimal. On the resulting reliable transitions, we introduce a *self-consistency loss* that enforces a state $\rightarrow$ image $\rightarrow$ state cycle on discriminator-gated imagination rollouts. Unlike prior cycle objectives limited to short, data-supported predictions, our self-consistency loss extends cycle supervision to multi-step, *policy-conditioned* imagination rollouts, extracting additional training signal specifically from states the agent is likely to visit and thereby promoting semantically meaningful state representations.

We evaluate AIDA across five MuJoCo tasks and two Gymnasium-Robotics tasks with scarce pre-collected target data, where it consistently outperforms baselines, demonstrating its effectiveness in practically important yet challenging regime. Our contributions are:

(i) **AIDA** A novel paradigm that augments limited target data with reliable synthetic transitions and leverages self-consistency to learn semantically meaningful representations.
(ii) **Discriminator-gated imagination.** A mechanism for generating only reliable imagined transitions by dynamically truncating imagination rollouts based on the discriminator's domain classification signal.
(iii) **Self-consistency via imagination-augmented data.** A scheme that learns semantically meaningful representations by enforcing a self-consistency loss on synthetic policy-conditioned imagination rollouts, providing additional supervision under limited target data.
(iv) **Experimental evaluation.** Comprehensive experiments across five MuJoCo tasks and two Gymnasium-Robotics tasks demonstrate that imagined transitions progressively become unreliable as the rollout horizon grows and that discriminator-gated truncation yields higher returns than any fixed-horizon alternative, enabling AIDA to consistently outperform baseline DA methods in the low-data regime.

2 Related Work

Bridging the sim-to-real gap has been a longstanding challenge in RL, and a number of approaches have been proposed to address the shift in state distributions between source and target domains. Broadly, these methods fall into three categories: domain randomization (DR), domain generalization (DG), and domain adaptation (DA).

Domain Randomization (DR). DR trains an agent in simulation under a wide range of environment parameters—such as textures, lighting, camera poses, and object appearances—with the goal that such visual variations will cover the target-domain state distribution. In vision-based settings, this typically involves randomizing camera poses, lighting conditions, object placements, textures (including non-realistic ones), and backgrounds [9, 10, 15, 18]. The resulting randomized images can be fed directly to the agent for policy learning,

or first passed through an image-translation generator to produce more realistic observations before being used for training.

However, DR often requires generating a large number of variants to encourage the policy to learn invariant representations, which makes training computationally expensive. More importantly, the randomization space (i.e., which factors to vary and their ranges) is manually specified by the practitioner; therefore, target domains may contain additional, unforeseen factors that are not included in the chosen variants. As a result, the assumed variants may fail to fully cover the target-domain distribution, leading to degraded transfer performance [6].

Domain Generalization (DG). While DR broadens the training distribution by randomizing simulator parameters to cover a particular target domain, DG assumes no target-domain data at training time and seeks to generalize to unseen domains by learning domain-invariant representations from diverse sources/augmentations. In vision-based RL, DG is commonly implemented by shaping the visual representation to be insensitive to appearance changes—for example, through visual augmentations [7], representation-level consistency and foreground extraction [21], or feature factorization/reconstruction objectives [3]. However, because the test domain is unknown, there is no guarantee that the training sources/augmentations capture the true factors of variation encountered at deployment; consequently, when the domain gap is large or includes unmodeled shifts, DG performance may degrade substantially.

Domain Adaptation (DA). Unlike DR and DG, DA assumes access to a target environment and adapts a source-trained policy to the target domain before deployment or at test time [2, 6, 16, 20]. Because the adaptation is driven by target-domain data, DA can reduce the source–target shift even when the primary factors behind the domain gap are unknown, focusing updates on discrepancies that are actually observed in the target domain.

Concretely, a common design is to adapt *what the policy sees* (i.e., the observation/feature representation) so that source and target become consistent, or to adapt the *policy itself* using target-domain experience. For example, PAD [6] performs test-time adaptation by updating the visual encoder with an inverse-dynamics self-supervised objective, while CODAS [2] uses a GAN-based mapping to align target visual observations with source-side privileged information. PRFT [20] instead adapts the policy during deployment by fine-tuning with predicted rewards computed in the target domain, enabling reward-free adaptation. Sun et al. [16] transfer source-trained latent transition and reward models as fixed regularizers to guide representation learning in the target domain, effectively encouraging target features to be compatible with source dynamics priors. Despite their effectiveness, these methods typically assume access to substantial target data, which is often impractical in real-world deployment where data collection is costly and limited. In contrast, our work explicitly targets the scarce-data regime, which remains underexplored despite its practical importance.

3 Preliminaries

3.1 Markov Decision Process.

We consider a Markov Decision Process defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p(s' \mid s, a)$ is the transition dynamics, $r(s, a)$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. The goal of RL is to learn a policy $\pi(a \mid s)$ that maximizes the expected cumulative return $\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right]$. In visual RL, the agent does not observe the underlying state s directly; instead, it receives a high-dimensional image observation $o \in \mathcal{O}$, and must learn to act from pixels.

A dynamics model $f_\omega : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is a parameterized approximation of the transition dynamics $p(s' \mid s, a)$. Given a state s_t and action a_t, it predicts the next state $\tilde{s}_{t+1} = f_\omega(s_t, a_t)$. We learn ω by minimizing a one-step prediction loss such as $\mathcal{L}_{wm} = \mathbb{E}_{(s,a,s') \sim p(s'|s,a)}\left[\|f_\omega(s, a) - s'\|_2^2\right]$. By repeatedly applying f_ω under a policy π, the agent generates multi-step rollouts $\tilde{s}_{t+1}, \tilde{s}_{t+2}, \ldots$ entirely in state space without interacting with the environment—a process referred to as *imagination* [5,17]. Since the dynamics model is an approximation, prediction error accumulates over successive steps, making longer rollouts less reliable.

3.2 Domain Adaptation in RL as Variational Inference.

When the source domain provides low-dimensional states while the target domain provides image observations, a mapping function must project target observations into the source state space to enable policy transfer. We build on the trajectory-level variational inference framework of [2] for aligning inferred and source state distributions. This framework models a *generation process* via an auto-regressive observation model $p_\theta(o_t \mid \hat{s}_t, o_{t-1})$ that renders target observations from inferred states $\hat{s}_t$ and the previous observation o_{t-1}, and an *inference process* via a mapping function $q_\phi(s_t \mid \hat{s}_{t-1}, a_{t-1}, o_t)$ that infers a mapped state $\hat{s}_t$ by combining the previous inferred state $\hat{s}_{t-1}$, the previous action a_{t-1}, and the target observation o_t. Applying step-wise inference over a full trajectory defines a variational posterior $q_\phi(\tau^s \mid \tau^o)$ over the source state–action trajectory $\tau^s = \{(s_1, a_1), \ldots, (s_T, a_T)\}$ given the target observation–action trajectory $\tau^o = \{(o_1, a_1), \ldots, (o_T, a_T)\}$. The resulting ELBO is:

$$\max_{\phi,\theta} \mathbb{E}_{\tau^o}\Big[\mathbb{E}_{\hat{\tau}^s \sim q_\phi(\tau^s|\tau^o)}\big[\log p_\theta(\tau^o \mid \hat{\tau}^s)\big] - D_{\mathrm{KL}}(q_\phi(\tau^s \mid \tau^o) \parallel p(\tau^s))\Big],$$

where $\hat{\tau}^s$ denotes an inferred state–action trajectory sampled from q_ϕ, the first term is a reconstruction loss $\mathcal{L}_{\text{recon}}$ enforcing visual consistency, and the second term aligns the inferred trajectory distribution with the source prior. Since the KL term is intractable, it is approximated with the adversarial loss [14]: a binary discriminator D_η distinguishes source trajectories from inferred ones, while q_ϕ acts as the generator. We denote this loss as $\mathcal{L}_{\text{adv}}$:

$$\min_{\phi} \max_{\eta} \mathbb{E}_{\tau^s \sim \mathcal{D}_{src}}[\log D_\eta(\tau^s)] + \mathbb{E}_{\tau^o \sim \mathcal{D}_{tgt},\, \hat{\tau}^s \sim q_\phi(\cdot|\tau^o)}\big[\log\big(1 - D_\eta(\hat{\tau}^s)\big)\big]. \quad (1)$$

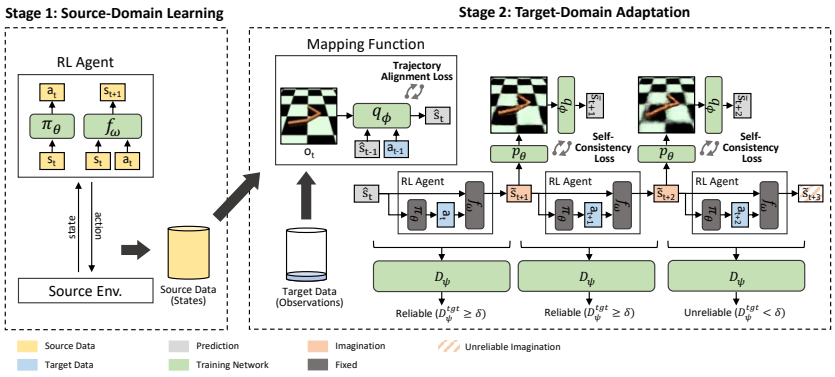


Fig. 1: Overview of AIDA. The framework proceeds in two stages. **Stage 1 (source-domain learning):** the agent interacts with the source environment to train a dynamics/world model f_ω and a policy π via dynamics learning and reinforcement learning. **Stage 2 (target-domain adaptation):** given scarce target image trajectories $\mathcal{D}_{\text{tgt}}$, we freeze f_ω and π and generate *reliable, adaptively truncated* imagined trajectories $\mathcal{D}_{\text{img}}$. A mapping function q_ϕ and an observation model p_θ are then optimized with a trajectory alignment loss $\mathcal{L}_{\text{align}}$ and a self-consistency loss $\mathcal{L}_{\text{sc}}$ (state $\rightarrow$ image $\rightarrow$ state) to learn target-aligned imagination under limited target data.

Together with the reconstruction loss, this defines the trajectory alignment loss $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{adv}}$, where α controls the relative weight of the adversarial alignment term. Trajectory-level alignment is essential because state-level matching alone is ill-posed: a target observation may be mapped to an incorrect but plausible source state that is indistinguishable without sequential context.

4 Method

AIDA operates in two stages. In the first stage, a state-based policy π and a dynamics model f_ω are trained in the source domain with full environment access. In the second stage, given only a small pre-collected dataset of target-domain image trajectories $\mathcal{D}_{tgt}$, AIDA adapts the source-trained policy to the target domain without any additional target interaction. The adaptation builds on the trajectory alignment framework described in Sec. 3, optimizing the trajectory alignment loss. However, the core challenge is that scarce target data provides insufficient coverage for alignment. AIDA addresses this through adaptive imagination: a learned dynamics model generates synthetic rollouts from target-inferred states, a three-way discriminator truncates these rollouts when they drift from the target manifold, and a self-consistency loss on the resulting reliable transitions provides additional supervision for learning semantically meaningful representations. Our overall workflow is summarized in Fig. 1. We begin by formally defining the problem setting (Sec. 4.1), then describe each

component in detail. Additional implementation details are provided in the supplementary material.

4.1 Problem Formulation

We consider domain adaptation for visual RL between a source and a target domain. We distinguish the underlying *state* $s \in \mathcal{S}$, which specifies the physical configuration of the system, from the *observation* received by the agent. Following common sim-to-real DA formulations [2, 6, 16], we assume that the source and target share the same state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics, and reward function, but differ in their observation spaces.

A straightforward approach to sim-to-real visual adaptation is *image-to-image adaptation*, which encodes source and target observations into a shared latent feature space and aligns the resulting representations so that a policy trained on the aligned features can transfer across domains [6, 16, 20]. However, visual-based RL is notorious for unstable training [11, 23] and expensive computation [22], making image-to-image approaches impractical when interaction and computation budgets are limited. Moreover, most existing domain adaptation methods assume access to a large amount of target-domain data [2], which is often unrealistic since collecting real-world data is far more costly than generating simulated data.

In this work, we consider a challenging setting that addresses both of these limitations. First, we adopt a *cross-modality* problem setting. The policy is trained in the source domain using low-dimensional simulator states $s \in \mathcal{S}$, while the target domain provides only high-dimensional image observations $o \in \mathcal{O}$. Simulators typically provide privileged access to the full state, which enables substantially more efficient policy learning. However, this design turns deployment into a cross-modality transfer problem (sim(state) $\rightarrow$ real(image)) and enlarges the domain gap compared to image-to-image adaptation, since the policy must operate from a different sensing modality at test time. Second, we consider a *scarce target data* regime. Only a small number of target-domain trajectories are available, and no additional target interaction is permitted during adaptation. This makes the adaptation problem substantially harder, as the alignment module must generalize from minimal target samples.

Consequently, domain adaptation in our setting reduces to learning an image-to-state mapping $q_\phi \colon \mathcal{O} \rightarrow \mathcal{S}$ from scarce target data, projecting target observations into the source state space and enabling direct deployment of the source policy in the target domain as $\pi(q_\phi(o))$. The target dataset $\mathcal{D}_{tgt}$ for training q_ϕ consists of trajectories of image observations paired with executed actions, *e.g.*, tuples (o_t, a_t, o_{t+1}), and contains no privileged target states.

4.2 Discriminator-Gated Imagination

When target data is scarce, the alignment objective in Eq. (1) alone provides insufficient supervision. To obtain additional training signal, we leverage the learned dynamics model to generate imagination rollouts from target-inferred

states, synthesizing auxiliary transitions beyond the limited target dataset. A key question, however, is how long to roll out. If the rollout horizon is too long, imagined transitions drift into regions unsupported by target data, where the mapping is unreliable. Additionally, the dynamics model accumulates prediction error over successive steps, further degrading rollout quality. If it is too short, too few synthetic transitions are generated to supplement the scarce target data. Therefore, effective adaptation requires choosing an imagination horizon that is long enough to provide useful supervision while remaining within regions where the imagined transitions are reliable.

To address this, we introduce a three-way discriminator D_ψ (distinct from the binary alignment discriminator D_η in Sec. 3.2) trained to classify transitions as (i) *target-inferred*, (ii) *source*, or (iii) *imagined*. Since the discriminator is trained on real target transitions, it assigns high target-likeness scores to imagined transitions that resemble the target data and low scores to those that have drifted away, making its output a natural indicator of rollout reliability. We use this score as an online confidence signal to adaptively truncate rollouts, continuing while the score remains high and stopping when it drops, so that only reliable transitions contribute to training. The three-way formulation is important because a binary *source-vs-target* discriminator scores transitions that drift outside the support of *both* domains ambiguously (often near 0.5), which may be mistakenly interpreted as "moderately target-like" even though they are unlike either distribution. Introducing an explicit *imagined* class allows the discriminator to recognize such out-of-support rollouts as *imagined/OOD* rather than assigning a misleading intermediate probability.

Utilizing the three-way discriminator as an online confidence signal, we now describe how imagination rollouts are generated and adaptively truncated during training. Given a target-initialized state $\hat{s}_t \sim q_\phi(s_t \mid \hat{s}_{t-1}, a_{t-1}, o_t)$, we roll out imagined states $\tilde{s}_{t+1}, \tilde{s}_{t+2}, \ldots$ using the dynamics model under the frozen policy π. Let $D_\psi^{\text{tgt}}(x) \triangleq P_\psi(y = \text{target-inferred} \mid x)$ denote the discriminator's predicted probability of the *target-inferred* class. At each step, we use $D_\psi^{\text{tgt}}(\cdot)$ as a *target-likeness* score and define the adaptive rollout horizon as the first step at which this score falls below a threshold $\delta \in (0, 1)$, limited to a maximum horizon $K_{\max}$,

$$K^* = \min\left(\min\left\{k \geq 0 \;\middle|\; D_\psi^{\text{tgt}}(\tilde{s}_{t+k},\, a_{t+k},\, \tilde{s}_{t+k+1}) < \delta\right\},\, K_{\max}\right), \tag{2}$$

so that imagination proceeds for steps $k = 0, 1, \ldots, K^* - 1$. A higher D_ψ^{tgt} indicates that the rollout remains within the target-aligned manifold, whereas a low score signals drift into a low-confidence region where self-consistency training would be counterproductive.

Discriminator-gated imagination offers more than simply finding an appropriate rollout length, because the truncation decision is made per-state rather than globally. This is crucial since each state from which imagination begins leads to a different rollout trajectory, and the maximum length of reliable imagination varies accordingly. In practice, we observe that the rollout horizon varies significantly depending on the state (see 5.2.Q2), confirming that a single fixed

K cannot capture this per-state variability, and that the adaptive scheme consistently outperforms all fixed-horizon alternatives (see 5.2.Q3).

4.3 Self-Consistency via Imagination-Augmented Data

Here we describe how to fully leverage the reliable imagined transitions obtained from discriminator-gated truncation. Our key idea is to impose a self-consistency constraint on these imagined transitions, requiring that a state decoded into an image and re-inferred should recover the original. This cycle-consistency idea has been explored in latent space [8, 25], but only on real data points. Our formulation applies this constraint to *imagined* states from multi-step rollouts, enabling supervision even in regions where no real target data exists. For each reliable imagined state $\tilde{s}_{t+k}$ $(k \leq K^*)$ admitted by the discriminator gate, we generate an imagined observation $\tilde{o}_{t+k} \sim p_\theta(\cdot \mid \tilde{s}_{t+k}, \tilde{o}_{t+k-1})$ and re-infer a state $\bar{s}_{t+k} \sim q_\phi(s_{t+k} \mid \tilde{s}_{t+k-1}, a_{t+k-1}, \tilde{o}_{t+k})$. We penalize discrepancies between the re-inferred and original imagined states,

$$\mathcal{L}_{\text{sc}} = \frac{1}{K^*} \sum_{k=1}^{K^*} \|\bar{s}_{t+k} - \tilde{s}_{t+k}\|_2^2 \,. \tag{3}$$

This loss encourages the encoder to capture semantically meaningful state information that is *visually grounded*. Factors that cannot be rendered into target-like observations and re-inferred are suppressed, while state-relevant factors that remain consistent under the decode→re-infer cycle are reinforced. Furthermore, since our consistency constraint is enforced over *multi-step* imagination rather than short, data-supported predictions, it yields extra supervision across successive rollout steps. The *policy-conditioned* rollouts further concentrate this supervision on *reachable* states the current policy is likely to encounter. Because reliable imagination can generate many such multi-step state–observation pairs without additional target interaction, the self-consistency objective provides dense auxiliary supervision beyond scarce real target data, stabilizing representation alignment in regions where real samples are insufficient.

Combining the trajectory alignment objective from Sec. 3 with the self-consistency loss, the overall training objective for Stage 2 is: $\mathcal{L} = \mathcal{L}_{align} + \lambda \mathcal{L}_{sc}$, where λ controls the relative weight of the self-consistency regularization.

5 Experiments

We design our experiments to investigate the following questions: (1) Does AIDA outperform existing domain adaptation baselines under scarce target data? (2) Does the three-way discriminator reliably detect unreliable drifted transitions? (3) Does the adaptive imagination outperform fixed-horizon imagination, where unreliable imagined transitions are inevitably included? (4) Does the self-consistency loss encourage the encoder to learn state representations that faithfully capture the true physical configuration? We first describe the experimental setup, then address each question in Sec. 5.2.

5.1 Experimental Setup

We evaluate AIDA on seven continuous control tasks: five tasks from the MuJoCo benchmark suite [19], HalfCheetah, Hopper, Swimmer, Walker2d, and InvertedPendulum, and two tasks from Gymnasium-Robotics, Shadow Dexterous Hand Reach and Fetch Reach. Following the problem setting in Sec. 4.1, the source domain provides low-dimensional states while the target domain provides only high-dimensional image observations, creating a large cross-modality gap. The target data budget is severely restricted to reflect the practical reality that real-world data collection is costly and limited: only **50 trajectories** are available for all tasks—1/6 of the target data used in previous work [2]. For InvertedPendulum, a relatively simple task, we reduce the trajectory length T and image resolution to avoid performance saturation across all methods. The target trajectories are collected by executing the expert policy in the target environment, and no additional interaction with the target environment is permitted during adaptation, except for PAD, the online adaptation baseline. The source-domain policy π is trained using Soft Actor-Critic (SAC) [4] with low-dimensional state inputs until convergence, then frozen throughout the adaptation procedure. Additional details are provided in the supplementary material.

We compare AIDA against five methods that span different approaches to bridging the observation mismatch:

- **CODAS** [2]: A GAN-based domain adaptation method that aligns target visual observations with source-side privileged state information via a variational inference framework with adversarial training. As the closest prior work to our setting, comparing against CODAS directly measures the benefit of adaptive imagination under the same problem setting.
- **GAN_STACK** [2]: A GAN-based domain adaptation method that stacks consecutive observations and feeds them to the alignment module, providing temporal context. This baseline tests whether simply providing more temporal information can substitute for the imagination-based augmentation that AIDA employs.
- **PAD** [6]: A self-supervised test-time adaptation method originally proposed for adapting vision-based RL policies to unseen visual changes. Unlike the other offline baselines, PAD continuously interacts with the target environment during deployment. This baseline evaluates whether self-supervised online adaptation is sufficient under our larger state-to-image domain gap, compared to AIDA.
- **Behavior Cloning (BC)**: A supervised baseline that directly learns a mapping from target image observations to source states using paired data, without any adversarial or self-consistency objective. This serves as a reference for how far pure supervised learning can go with scarce paired data.
- **Oracle (Image)**: SAC combined with a convolutional autoencoder [23] trained jointly with the RL objective on target-domain image observations with *unlimited* online interaction. This serves as an approximate upper bound for visual RL without cross-modal transfer, providing a reference point for the offline adaptation methods.

Table 1: Comparison of domain adaptation methods under scarce target data. RMSE (↓) and Return Ratio (↑) report converged performance. A ratio of 1.0 corresponds to the source expert's average return. Best results are in **bold**. Mean±std over 3 seeds.

(a) Return Ratio (↑)

Method	H.Cheetah	Hopper	Swimmer	Walker2d	Inv.Pend.	Shadow	Fetch
Oracle (image)	1.638±.343	0.591±.232	0.304±.060	0.196±.073	0.989±.072	0.991±.045	0.975±.076
BC	0.462±.011	0.121±.018	0.403±.042	**0.101±.013**	0.319±.052	0.836±.033	0.453±.009
GAN_STACK	0.340±.087	0.056±.008	0.261±.066	0.026±.004	0.063±.004	0.645±.122	0.823±.053
PAD	0.573±.075	0.142±.023	0.347±.058	0.036±.005	0.155±.012	0.395±.088	0.821±.044
CODAS	0.711±.098	0.356±.086	0.468±.069	0.038±.009	0.440±.038	0.893±.059	0.897±.007
AIDA (ours)	**0.810±.113**	**0.440±.142**	**0.512±.047**	0.058±.018	**0.561±.049**	**0.931±.027**	**0.984±.012**

(b) RMSE (↓)

Method	H.Cheetah	Hopper	Swimmer	Walker2d	Inv.Pend.	Shadow	Fetch
GAN_STACK	3.006±.153	0.862±.059	1.360±.272	2.525±.317	0.328±.042	0.452±.029	0.025±.009
CODAS	1.550±.081	0.540±.021	0.482±.057	2.180±.405	0.108±.007	0.425±.037	0.017±.004
AIDA (ours)	**1.403±.116**	**0.523±.023**	**0.356±.015**	**1.784±.320**	**0.075±.004**	**0.391±.016**	**0.015±.008**

Evaluation metrics. We report two complementary metrics:

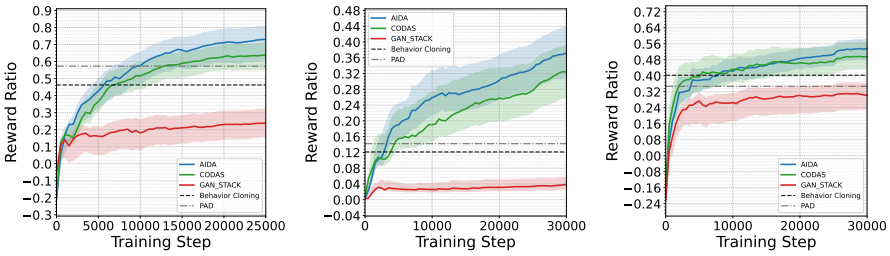
- **Root Mean Squared Error (RMSE)**: Measures the accuracy of the learned mapping q_ϕ by computing the root-mean-squared distance between the inferred states and the ground-truth states.
- **Return Ratio**: Measures the task performance of the adapted policy relative to the source-trained expert. Defined as $r_{\text{ratio}} = r/r^*$, where r is the cumulative return of the adapted policy deployed in the target environment, and r^* is the average return of the source expert. A ratio of 1.0 indicates full recovery of the source expert's performance.

5.2 Experimental Results and Analysis

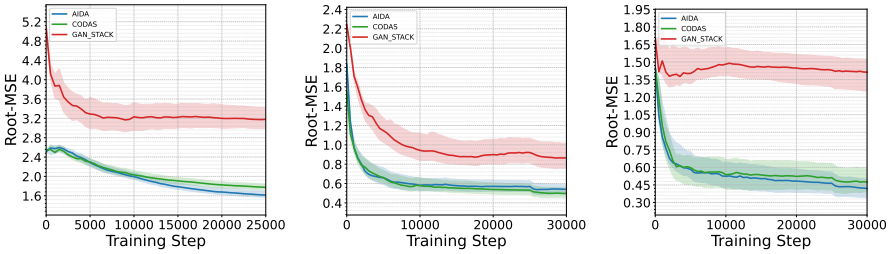
(Q1) Does AIDA outperform existing domain adaptation baselines under scarce target data? To assess whether adaptive imagination provides effective additional supervision under scarce target data, we compare AIDA against all baselines across the five MuJoCo tasks and two Gymnasium-Robotics tasks. Note that RMSE directly measures how accurately q_ϕ recovers the true proprioceptive state, serving as a quantitative indicator of representation quality.

Table 1 summarizes the results. AIDA achieves the lowest RMSE on every task and the highest return ratio on six out of seven tasks. As shown in Fig. 2, AIDA also converges faster than all baselines, with the RMSE curves showing a consistent trend of steeper decrease and stabilization at a lower value, though the difference on Hopper is marginal.

Comparison with PAD further highlights AIDA's robustness to large domain gaps. Although PAD continuously interacts with the target environment during deployment, it shows lower performance than AIDA across all tasks. This



(a) Return – HalfCheetah **(b)** Return – Hopper **(c)** Return – Swimmer



(d) RMSE – HalfCheetah **(e)** RMSE – Hopper **(f)** RMSE – Swimmer

Fig. 2: Training curves across three environments. Top row: target-domain return (↑) over training iterations. Bottom row: mapping RMSE (↓) over training iterations. Solid lines denote the median and shaded regions represent the interquartile range over 3 random seeds.

suggests that while PAD's self-supervised adaptation can handle relatively small image-to-image domain gaps, AIDA's adaptation strategy is more effective under the larger state-to-image domain gap considered in our setting.

Notably, on Swimmer and Fetch, AIDA even outperforms Oracle (image), which has access to unlimited online interaction. This is because AIDA benefits from a policy stably trained on low-dimensional states, whereas Oracle (image) must learn both representations and a policy jointly from images. Oracle (image) also achieves relatively low return ratios on Swimmer (0.304) and Walker2d (0.196), suggesting that these tasks are inherently challenging for extracting meaningful state information from images alone. On Walker2d, all DA methods yield lower return ratios than BC, despite AIDA achieving the best RMSE. We attribute this to the balance-critical nature of Walker2d: even small mapping errors can cause the agent to fall, amplifying the gap between reconstruction accuracy and downstream performance.

Overall, AIDA's advantage stems from two factors. The adaptive imagination mechanism augments the scarce target data with reliable synthetic transitions, and the self-consistency loss leverages these transitions to learn semantically meaningful representations. Together, they provide dense supervision that existing methods, which rely solely on the limited real target data, cannot achieve.

(Q2) Does the three-way discriminator reliably detect unreliable drifted transitions? The adaptive truncation mechanism relies on the three-way dis-

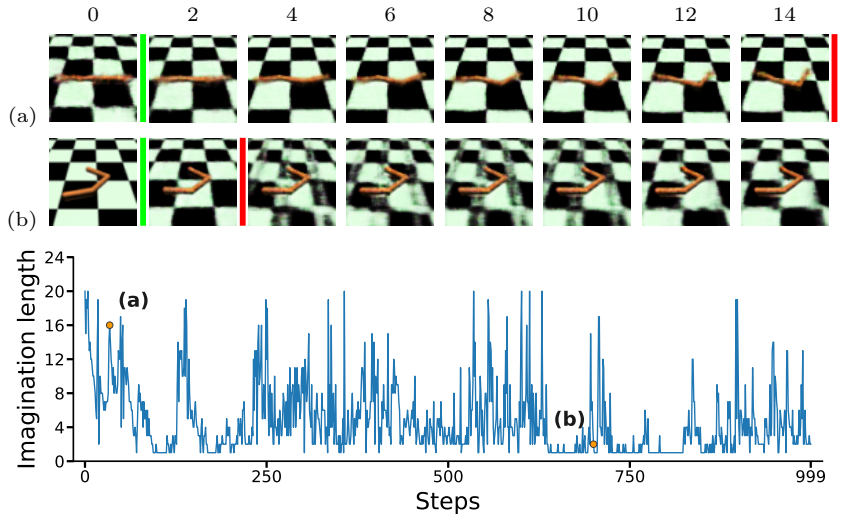


Fig. 3: Top: imagined observations decoded from adaptive imagination rollouts on the Swimmer task (shown at every 2 steps). The green line marks the point at which imagination begins. The red line indicates the truncation point determined by D_ψ. Bottom: adaptive imagination horizon at each step within a single episode, illustrating how the truncation length varies depending on the state encountered at each step. The two highlighted rollouts, **(a)** and **(b)**, correspond to the marked points in the bottom plot (orange dots) and show the decoded imagination sequences starting from those specific steps.

criminator D_ψ to detect when imagined transitions drift away from the target-aligned manifold. To verify whether D_ψ makes reasonable truncation decisions, we visualize imagined observations decoded from rollouts across two initial states, and additionally plot the adaptive imagination horizon at each step within a single episode alongside the corresponding observations at selected steps.

As shown in the top of Fig. 3, the decoded observations before the truncation point remain visually coherent and physically plausible, whereas those after the truncation point exhibit noticeable degradation such as distorted body configurations and blurred artifacts. This clear quality gap confirms that the discriminator reliably identifies the boundary between reliable and drifted imagination. Notably, the truncation length varies substantially across states. The bottom of Fig. 3 visualizes this per-state variability by plotting the adaptive imagination horizon across steps within a single episode, with the marked points corresponding to the two example rollouts shown above. The horizon fluctuates significantly depending on the encountered state, confirming that imagination reliability is highly state-dependent and motivating adaptive, per-transition truncation rather than a single globally fixed horizon.

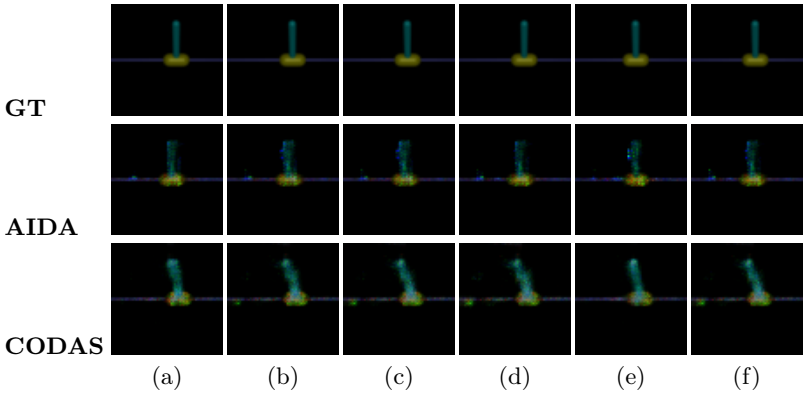


Fig. 5: Qualitative comparison of imagined observations on the InvertedPendulum task. Each column shows the ground-truth observation (GT) and those decoded by AIDA (with $\mathcal{L}_{sc}$) and CODAS (without $\mathcal{L}_{sc}$) for the same state.

(Q3) Does the adaptive imagination outperform fixed-horizon imagination, where unreliable imagined transitions are inevitably included? The qualitative analysis in Fig. 3 revealed that imagination reliability varies significantly across individual states and that the discriminator can detect this drift. We now examine whether including these unreliable transitions actually harms performance, by comparing the adaptive horizon against fixed horizons $K \in \{0, 5, 10, 20\}$.

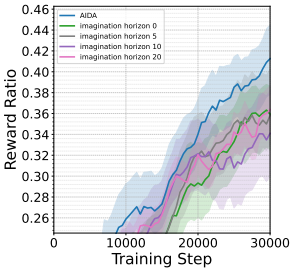


Fig. 4: Adaptive vs. fixed-K training curves on Hopper. Return (↑). Mean over 3 seeds.

As shown in Fig. 4, AIDA achieves the highest return among all configurations on Hopper. Fixed horizons—regardless of length—yield performance comparable to or even worse than no imagination at all, confirming that indiscriminately including unreliable transitions harms rather than helps adaptation. In contrast, the adaptive scheme selectively retains only high-confidence transitions, providing clean supervision that enables meaningful representation learning beyond what scarce target data alone can offer.

(Q4) Does the self-consistency loss encourage the encoder to learn state representations that faithfully capture the true physical configuration? To examine whether the self-consistency loss encourages semantically meaningful state representations, we compare the quality of imagined observations from AIDA and CODAS. Since the observation model p_θ can only render realistic images from states that accurately reflect the true physical configuration, the quality of decoded images directly indicates how well the learned state representations capture the actual physical state. We select InvertedPendulum

for this analysis because its reduced trajectory length and image resolution make the difference between methods more pronounced.

As shown in Fig. 5, the challenging setting leads to lower reconstruction quality overall, which makes the outputs of AIDA and CODAS appear visually similar. A closer comparison, however, reveals a clear difference. AIDA closely matches the ground-truth pendulum angle and position across all cases, whereas CODAS consistently depicts angles that deviate from the true configuration. This suggests that without the self-consistency loss, the learned state representations may not fully capture the precise physical configuration even when the decoded images appear plausible. The self-consistency loss helps q_ϕ learn semantically meaningful state representations that more faithfully reflect the true pendulum orientation.

6 Conclusions

We presented AIDA, a domain adaptation framework for visual reinforcement learning that leverages adaptive imagination to overcome the challenge of limited target-domain data. AIDA augments scarce target data with reliable and semantically meaningful imagination rollouts. To ensure the reliability of these transitions, a three-way discriminator adaptively truncates drifted rollouts, and a self-consistency loss is trained on the resulting reliable transitions to learn semantically meaningful representations. Experiments on five MuJoCo tasks and two Gymnasium-Robotics tasks demonstrate that AIDA consistently outperforms existing baselines under severely limited target data, the discriminator reliably detects imagination drift, and adaptive truncation outperforms all fixed-horizon alternatives. Despite these results under a challenging cross-modality setting with scarce target data, our work has two limitations. First, the adapted policy still falls short of full source-expert performance, particularly on balance-critical tasks. Incorporating online adaptation algorithm could help close this gap. Second, AIDA assumes shared transition dynamics across domains, which may not hold in practice. Extending AIDA to handle dynamics mismatch is a promising direction for future work.

Acknowledgements

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program-Technology Innovation Program) (RS-2025-25449157, Development of Commercialization Technologies for End-to-End Autonomous Driving Products) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25497111), the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2026-RS-2023-00255968) grant funded by the Korea government (MSIT), and the Ajou University research fund.

References

1. Chen, J., Li, S.E., Tomizuka, M.: Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems **23**(6), 5068–5078 (2021)
2. Chen, X.H., Jiang, S., Xu, F., Zhang, Z., Yu, Y.: Cross-modal domain adaptation for cost-efficient visual reinforcement learning. Advances in Neural Information Processing Systems **34**, 12520–12532 (2021)
3. Choi, H., Lee, H., Jeong, S., Min, D.: Environment agnostic representation for visual reinforcement learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 263–273 (2023)
4. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: 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)
5. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations (2020)
6. Hansen, N., Jangir, R., Sun, Y., Alenyà, G., Abbeel, P., Efros, A.A., Pinto, L., Wang, X.: Self-supervised policy adaptation during deployment. In: International Conference on Learning Representations (2021)
7. Hansen, N., Wang, X.: Generalization in reinforcement learning by soft data augmentation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13611–13617. IEEE (2021)
8. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
9. James, S., Davison, A.J., Johns, E.: Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In: Conference on Robot Learning. pp. 334–343. PMLR (2017)
10. James, S., Wohlhart, P., Kalakrishnan, M., Kalashnikov, D., Irpan, A., Ibarz, J., Levine, S., Hadsell, R., Bousmalis, K.: Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12627–12637 (2019)
11. Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al.: Model-based reinforcement learning for atari. In: International Conference on Learning Representations (2020)
12. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
13. Niu, H., Chen, Q., Liu, T., Li, J., Zhou, G., ZHANG, Y., HU, J., Zhan, X.: xted: Cross-domain adaptation via diffusion-based trajectory editing. In: NeurIPS 2024 Workshop on Open-World Agents (2024)
14. Nowozin, S., Cseke, B., Tomioka, R.: f-gan: Training generative neural samplers using variational divergence minimization. Advances in neural information processing systems **29** (2016)
15. Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., Abbeel, P.: Asymmetric actor critic for image-based robot learning. In: Robotics: Science and Systems (2018)

16. Sun, Y., Zheng, R., Wang, X., Cohen, A., Huang, F.: Transfer rl across observation feature spaces via model-based regularization. In: International Conference on Learning Representations (2022)
17. Sutton, R.S.: Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin **2**(4), 160–163 (1991)
18. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 23–30. IEEE (2017)
19. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)
20. Wang, W., Fang, X., Hager, G.: Adapting image-based rl policies via predicted rewards. In: 6th Annual Learning for Dynamics & Control Conference. pp. 324–336. PMLR (2024)
21. Wang, X., Lian, L., Yu, S.X.: Unsupervised visual attention and invariance for reinforcement learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6677–6687 (2021)
22. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)
23. Yarats, D., Zhang, A., Kostrikov, I., Amos, B., Pineau, J., Fergus, R.: Improving sample efficiency in model-free reinforcement learning from images. In: Proceedings of the aaai conference on artificial intelligence. vol. 35, pp. 10674–10681 (2021)
24. Zhao, W., Queralta, J.P., Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: 2020 IEEE symposium series on computational intelligence (SSCI). pp. 737–744. IEEE (2020)
25. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in neural information processing systems **30** (2017)