Overcoming Knowledge Barriers: Online Imitation Learning from Observation with Pretrained World Models

Xingyuan Zhang
xingyuan.zhang@volkswagen.de
Volkswagen AG
Technical University of Munich

Philip Becker-Ehmck
philip.becker-ehmck@volkswagen.de
Volkswagen AG

Patrick van der Smagt
smagt@volkswagen.de
Volkswagen AG
Eötvös Loránd University Budapest

Maximilian Karl
maximilian.karl@volkswagen.de
Volkswagen AG

Abstract

Pretraining and finetuning models has become increasingly popular. But there are still serious impediments in Imitation Learning from Observation (ILfO) with pretrained models. This study identifies two primary obstacles: the Embodiment Knowledge Barrier (EKB) and the Demonstration Knowledge Barrier (DKB). The EKB emerges due to the pretrained models' limitations in handling novel observations, which leads to inaccurate action inference. Conversely, the DKB stems from the reliance on limited demonstration datasets, restricting the model's adaptability across diverse scenarios. We propose separate solutions to overcome each barrier and apply them to Action Inference by Maximising Evidence (AIME), a state-of-the-art algorithm. This new algorithm, AIME-NoB, integrates online interactions and a data-driven regulariser to mitigate the EKB. Additionally, it uses a surrogate reward function to broaden the policy's applicability, addressing the DKB. Our experiments on tasks from the DeepMind Control Suite and Meta-World benchmarks show that AIME-NoB significantly enhances sample efficiency and performance, presenting a robust framework for overcoming the challenges in ILfO with pretrained models.

1 Introduction

We have been going through a paradigm shift from learning from scratch to pretraining and finetuning, in particular in Computer Vision (CV) (He et al., 2016; Radford et al., 2021; He et al., 2022) and Natural Language Processing (NLP) (Devlin et al., 2019; Radford et al.; Ouyang et al., 2022; Touvron et al., 2023a;b) fields due to the increasing availability of foundation models (Bommasani et al., 2021) and ever-growing datasets. However, it is still unclear how to adapt this new paradigm into decision-making, in particular what type of models we need to pretrain and how these models can be adapted to solve downstream tasks. Recent work (Zhang et al., 2023; DeMoss et al., 2023; Sekar et al., 2020; Rajeswar et al., 2023; Hansen et al., 2023a) showed that pretrained latent space world models enable successful and efficient transfer to new tasks with either reinforcement learning (Sekar et al., 2020; Rajeswar et al., 2023; Hansen et al., 2023a) or Imitation Learning from Observation (ILfO) (Zhang et al., 2023; DeMoss et al., 2023). ILfO (Torabi et al., 2018; 2019; Baker et al., 2022; Zhang et al., 2023; DeMoss et al., 2023; Liu et al., 2022a), especially from videos (Baker et al., 2022; Zhang et al., 2023; Liu et al., 2022a; DeMoss et al., 2023), is a more promising approach in this new paradigm since it does not require a handcrafted reward function.

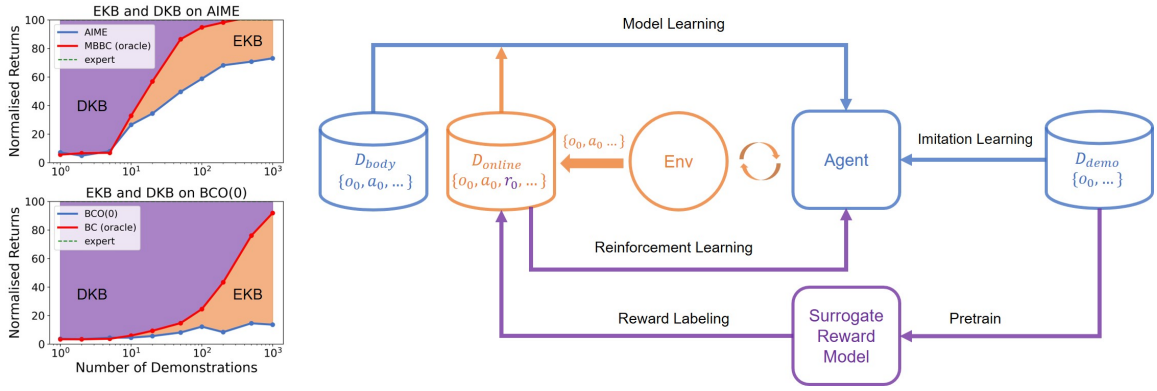


Figure 1: Main idea of this paper. On the left, we plot the performance of BCO(0) and AIME together with their oracle versions w.r.t. different number of demonstrations on walker-run task. The purple region between the oracle version and the expert is the Demonstration Knowledge Barrier (DKB) while the orange region between the algorithm and its oracle version represents the Embodiment Knowledge Barrier (EKB). On the right, we present the solutions proposed in this paper to overcome the two barriers. The blue parts represent the original version of the algorithms that suffer from the knowledge barriers. Orange parts demonstrate the solution for EKB, where the agent is allowed to interact with the environment and use D_{online} together with D_{body} to update the world model. Purple parts show the solution for DKB, where a surrogate reward model is pretrained and used to label the online dataset D_{online} and then used as an RL signal for policy learning.

But there are challenges when using pretrained models in ILfO. To quantify these we introduce two new barriers, which we call the Embodiment Knowledge Barrier (EKB) and the Demonstration Knowledge Barrier (DKB). The EKB describes the shortcomings of a pretrained model when confronted with novel observations and actions beyond its training experience. The DKB describes the generalisation from a limited number of expert demonstrations in imitation learning (Ho & Ermon, 2016). Approaches like BCO(0) (Torabi et al., 2018) and AIME (Zhang et al., 2023) typically suffer from these two knowledge barriers. First, these algorithms depend on the pretrained model to infer missing actions from observation sequences. Thus, when the model has not seen a specific observation before, it may not know enough about the embodiment to infer the correct action. Second, if the policy optimisation is only guided by limited demonstrations, it can lead to a policy that generalises poorly, working well in some scenarios but not in others.

To better showcase the two barriers, in Figure 1 left, we evaluate both AIME and BCO(0) and their oracle versions w.r.t. different number of demonstrations on walker-run task. Both algorithms pretrain a model from a large embodiment dataset and use that to infer the actions for the observation-only demonstrations. The oracle versions remove the need to infer the missing actions, thus removing the EKB. As we can see from the figure, the two algorithms are always upper-bounded by the corresponding oracle version, and the difference between them represents the EKB. On the other hand, even if given the true actions of the expert, imitation performance may still be impacted by a limited number of demonstrations providing insufficient coverage of the state space. Thus, the difference between the oracle version and the expert performance represents the DKB.

In this paper, we study how to resolve these barriers to improve the performance of ILfO approaches, in particular of AIME. For the EKB, we extend the setting from offline to online by allowing the agent to gather more data to train the world model. While for the DKB, we introduce a surrogate reward function to allow the policy to essentially train on more data. We demonstrate that the proposed modifications significantly improve the performance on nine tasks in DeepMind Control Suite (DMC) (Tunyasuvunakool et al., 2020) and six tasks in Meta-World (Yu et al., 2021).

2 Preliminary

We mostly follow the problem setup as described in Zhang et al. (2023). We consider a POMDP problem defined by the tuple $\{S, A, T, R, O, \Omega\}$, where S is the state space, A is the action space, $T : S \times A \to S$ is the dynamic function, $R : S \to \mathbb{R}$ is the reward function, O is the observation space, which is image in this paper, and $\Omega : S \to O$ is the emission function. The goal is to find a policy $\pi : S \to A$ which maximises the accumulated reward, i.e. $\sum_t r_t$.

We presume the existence of three datasets of the same embodiment available to our agent. The *embodiment dataset* D_{body} contains trajectories $\{o_0, a_0, o_1, a_1 \dots\}$ that represent past experiences of interacting with the environment. This dataset provides information about the embodiment for the algorithm to learn a world model. In addition, we also allow the agent to interact with the environment to collect new data in a *replay buffer* D_{online}. Note that, although the simulator will give us the reward information, the agent is not allowed to use them, and we only use the reward for evaluation purpose. The *demonstration dataset* D_{demo} contains a few expert trajectories $\{o_0, o_1, o_2 \dots\}$ of the embodiment solving a certain task defined by R_{demo}. The crucial difference between this dataset and the other two datasets is that the actions are not provided anymore since they are not observable from a third-person perspective. The goal of our agent is to learn a policy π from D_{demo} which can solve the task defined by R_{demo} as well as the expert who generated D_{demo}.

2.1 World Models

A World Model (Ha & Schmidhuber, 2018) is a generative model which models a probability distribution over sequences of observations, i.e. $p(o_{1:T})$. The model can be either unconditioned or conditioned on other factors, such as previous observations or actions. When the actions taken are known, they can be considered as the condition, i.e. $p(o_{1:T}|a_{0:T-1})$.

In this paper, we consider variational latent world models where the observation is governed by a Markovian hidden state. In the literature, this type of model is also referred to as a State-Space Model (SSM) (Karl et al., 2017; Hafner et al., 2019b;a; Becker-Ehmck et al., 2019; Klushyn et al., 2021). Such a variational latent world model involves four components, namely

$$\begin{aligned} &\text{encoder } z_t = f_\phi(o_t), && \text{posterior } s_t \sim q_\phi(s_t|s_{t-1}, a_{t-1}, z_t), \\ &\text{decoder } o_t \sim p_\theta(o_t|s_t), && \text{prior } s_t \sim p_\theta(s_t|s_{t-1}, a_{t-1}). \end{aligned}$$

$f_\phi(o_t)$ is the encoder to extract the features from the observation; $q_\phi(s_t|s_{t-1}, a_{t-1}, z_t)$ and $p_\theta(s_t|s_{t-1}, a_{t-1})$ are the posterior and the prior of the latent state variable; while $p_\theta(o_t|s_t)$ is the decoder that decodes the observation distribution from the state. ϕ and θ represent the parameters of the inference model and the generative model respectively.

Typically, the model is trained by maximising the Evidence Lower Bound (ELBO) which is a lower bound of the log-likelihood, or evidence, of the observation sequence, i.e. $\log p_\theta(o_{1:T}|a_{0:T-1})$. Given a sequence of observations, actions, and states, the objective function can be computed as

$$\text{ELBO} = \sum_{t=1}^{T} J_t^{\text{rec}} - J_t^{\text{KL}} = \sum_{t=1}^{T} \log p_\theta(o_t|s_t) - D_{\text{KL}}[q_\phi||p_\theta]. \tag{1}$$

The objective function is composed of two terms: the first term J^{rec} is the likelihood of the observation under the inferred state, which is usually called the reconstruction loss; while the second term J^{KL} is the KL divergence between the posterior and the prior distributions of the latent state. To compute the objective function, we use the re-parameterisation trick (Kingma & Welling, 2022; Rezende et al., 2014) to autoregressively sample the inferred states from the observation and action sequence. In summary, a world model is trained by solving the optimisation problem as

$$\phi^*, \theta^* = \underset{\phi,\theta}{\text{argmax}}\, \mathbb{E}_{\{o,a\}\sim D_{\text{body}}, s\sim q_\phi}[\text{ELBO}]. \tag{2}$$

2.2 AIME

AIME is a recently proposed algorithm that uses a pretrained world model to solve ILfO in an offline setting. Specifically, it uses the pretrained world model as an implicit inference model by solving for the best action sequence that makes the demonstration most likely under the trained world model. The imitation can be done jointly with the action inference using amortised inference and the re-parameterisation trick by solving the following optimisation problem

$$\psi^* = \underset{\psi}{\operatorname{argmax}}\, \mathbb{E}_{o\sim D_{\text{demo}}, s\sim q_{\phi^*,\theta^*}, a\sim\pi_\psi}[\text{ELBO}], \tag{3}$$

where ψ is the parameter for policy $\pi_\psi(a_t|s_t)$. The resulting objective is very similar to Equation (2), but with a subtle difference in the sampling path: observations are sampled from the dataset, while states and actions are iteratively sampled from the learned model and policy, respectively.

3 Methodology

In this section we will analyse the EKB and DKB for AIME. Based on the analysis we introduce a solution for each knowledge barrier and combine them into AIME-NoB, where NoB stands for **No B**arriers. The general framework of the solutions is shown in Figure 1 and the pseudocode of AIME-NoB is in Algorithm 1 in Appendix E.

3.1 Resolving the EKB

The most natural way to solve the EKB is to allow the agent to further interact with the environment. New experiences can minimise the error in the pretrained model in proximity of the policy π_ψ and gain more embodiment knowledge relevant for the task at hand. Torabi et al. (2018) proposed a modified version of BCO(0) called BCO(α) which introduced such an interaction phase. However, from their and our empirical results, it did not resolve the EKB since there remains a gap between BCO(α) and the BC oracle when the environment is complex. Continuing training an actor-critic from offline to online phases, as seen in recent works in Offline RL, also presents challenges, particularly regarding combating objective shifts (Lee et al., 2022; Ball et al., 2023; Nakamoto et al., 2023). Similarly, extending AIME from offline to online encounters issues like overfitting to newly collected datasets.

As the training progresses iteratively between data collection, model training and policy training, in the early phase of training there are only a few new trajectories available for training the model. Because the world model is highly expressive, it may overly favour similar trajectories in the new data, leading to a high ELBO. Normally, this may not be a big problem since, eventually, more and more data will be collected to combat this overfitting. But since AIME also depends on the ELBO to train the policy, it quickly causes the policy training to diverge.

In order to address the overfitting issue, we need a regulariser for model learning. Instead of using ad-hoc methods in the parameter space, we adopt a data-driven approach. Overfitting occurs due to a sudden shift from a large, diverse pretraining dataset to a small, narrow replay buffer. Appending the pretraining dataset to the replay buffer smooths this transition but reduces data efficiency. Instead, we sample separately from both datasets and modify the objective in Equation (2) to:

$$\phi^*, \theta^* = \underset{\phi,\theta}{\operatorname{argmax}}\, \alpha\mathbb{E}_{\{o,a\}\sim D_{\text{body}}, s\sim q_\phi}[\text{ELBO}] + (1-\alpha)\mathbb{E}_{\{o,a\}\sim D_{\text{online}}, s\sim q_\phi}[\text{ELBO}]. \tag{4}$$

The amount of data we sample from the pretraining dataset is controlled by a hyper-parameter α, which represents how much regularisation we put upon the model. Here we mainly consider setting $\alpha = 0.5$, so that we sample the data evenly from both datasets.

This finding contradicts Rajeswar et al. (2023) and Hansen et al. (2023a), where the pretrained world models do not need such a data-driven regulariser. We conjecture that unlike AIME, these approaches mainly use their world models purely as generative models to predict states and rewards given action sequences, which is only indirectly influenced by overfitting the ELBO.

3.2 Resolving the DKB

The straightforward way of solving the DKB is also to increase the number of demonstrations available to the agent. However, expert demonstrations are difficult and expensive to collect. Increasing the size of the demonstration dataset is not always feasible in real-world applications. In order to propose a more practical solution, we need to look deeper into what is the real cause of the DKB.

The policy-learning part of the AIME algorithm is essentially behaviour cloning, conducted only on the demonstration dataset. This provides clear guidance for covered states, but behavior is undefined for other states. AIME solely relies on the generalisation abilities of the learned latent state and the trained policy network to extrapolate the correct behaviour. In particular for small demonstration datasets, this can be unreliable or even impossible. Therefore, if we can enlarge the space of the covered states, we should reduce the DKB (Ross et al., 2011).

Based on this insight, we propose to introduce a surrogate reward providing guidance signal for the agent on the replay buffer dataset. Due to the instability of adversarial training (Goodfellow et al., 2014; Arjovsky et al., 2017; Ho & Ermon, 2016) and our focus on the pretraining paradigm, we opt to adopt the VIPER algorithm (Escontrela et al., 2023). Instead of training a discriminator, VIPER trains a video prediction model on the demonstration datasets and treats the likelihood of the video prediction model as the reward for policy learning, i.e. $r_t^{\text{VIPER}} = \log p_\nu(o_t|o_{<t})$.

Using this reward, we train the policy with an actor-critic algorithm based on imagination in the latent space of the world model (Hafner et al., 2019a). In order to do this, we first need to modify the reconstruction term in Equation (1) by adding an extra term for decoding the VIPER reward, i.e. $\log p_\theta(r_t^{\text{VIPER}}|s_t)$. Then, we further train a value estimator $V_\xi^\lambda(s_t)$ using TD(λ)-return estimates. Details of derivation can be found in Appendix C. Using this value estimate, we extend the policy objective of Equation (3) to

$$\psi^* = \operatorname*{argmax}_\psi \mathbb{E}_{o\sim D_{\text{demo}}, s\sim q_{\phi,\theta}, a\sim\pi_\psi}[\text{ELBO}] + \beta\mathbb{E}_{\{o,a\}\sim D_{\text{online}}, s\sim q_\phi, a'\sim\pi_\psi, s'\sim p_\theta}[V_{\xi'}^\lambda(s')], \tag{5}$$

where β is a hyper-parameter for balancing the two terms. We set $\beta = 0.1$ by default based on the difference of default learning rate in AIME and Dreamer.

4 Experiments

We aim to answer the following questions: a) How does the proposed AIME-NoB compare with state-of-the-art methods on common benchmarks? b) How well does the proposed modification resolve the EKB and the DKB? c) How do different choices of hyper-parameters influence the results? In order to answer these questions, we design our experiments on DMC and Meta-World benchmarks.

4.1 Datasets

For the DMC benchmark, we choose nine tasks across six embodiments following Liu et al. (2022a) and use their published dataset as the demonstration datasets. Each dataset contains only 10 trajectories to reflect the scarcity of expert demonstrations. For the embodiment dataset, in order not to leak the task information from the pretraining phase, we follow Rajeswar et al. (2023) and run a Plan2Explore (Sekar et al., 2020) agent for each embodiment with 2M environments steps and use its replay buffer as the embodiment dataset.

For Meta-World benchmark, we use the data and model from Hansen et al. (2023a). The embodiment dataset was created from the replay buffer datasets. The open-sourced replay buffer datasets contain 40k trajectories for each of the 50 tasks with only state information. In order to fit to our image observation setup, we render the images by resetting the environment to the initial state of each trajectory and then executing the action sequence. The details can be found in Appendix H.

Following the idea of not leaking too much about the task information, inspired by the common practice in offline RL benchmarks (Fu et al., 2021), we use the first 200 trajectories from each

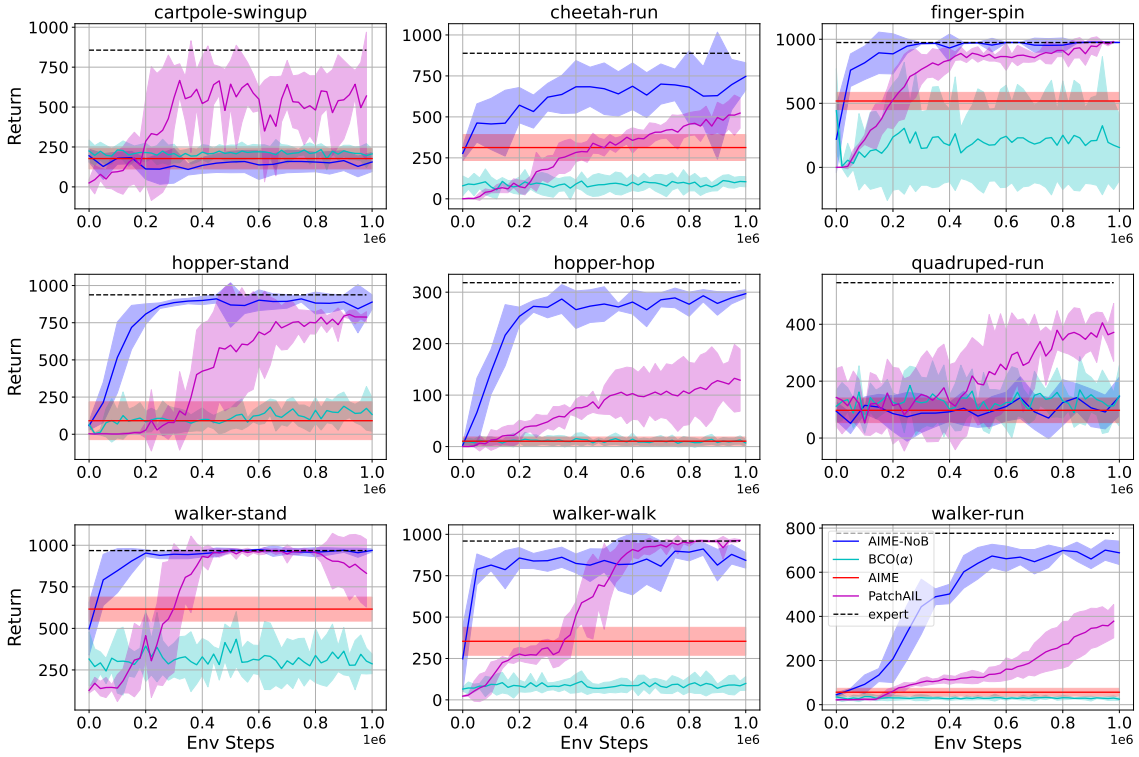


Figure 2: Benchmark results and 9 DMC tasks. Return are calculated by running the policy 10 times with the environment and taking the average return. The results are averaged across 5 seeds with the shade region representing 95% CI.

replay buffer and form a dataset with 10k trajectories in total. We call this dataset *MW-mt50*. To further study the out-of-distribution transfer ability of the pretrained model, we follow the difficulty classification of the tasks from Seo et al. (2022a) and only use the 39 easy and medium difficulty tasks to generate the datasets and the 11 tasks hard and very hard tasks as hold-out tasks. We uniformly sample 250 trajectories from the first 10k trajectories from each of the 39 tasks and form a dataset with 9750 trajectories in total. We refer to this dataset as *MW-mt39*. Hence, *MW-mt39* contains some expert behaviour solving the tasks, while *MW-mt50* consists of mostly exploratory behaviour.

As the demonstration datasets, we use the single-task policies open-sourced by TD-MPC2 and collect 50 trajectories for each tasks. We ensure that every trajectory in the demonstration dataset is successful. Since there are 500 steps in a DMC trajectory and only 100 steps in a Meta-World trajectory, the resulting datasets are roughly the same size.

4.2 Implementation

For the world model, we use the RSSM architecture (Hafner et al., 2019b) with the hyper-parameters in Hafner et al. (2019a) for DMC tasks. In addition, we use the KL Balancing trick from Hafner et al. (2020) to make the training more stable. For Meta-World, since the visual scene is more complex, we use the M size model from Hafner et al. (2023), but still with the continuous latent variable to be aligned with other models used in this paper. The policy network is implemented with a two-layer MLP, with 128 neurons for each hidden layer. All the models are trained with Adam optimiser (Kingma & Ba, 2017). More details about the hyper-parameters can be found in Appendix F.

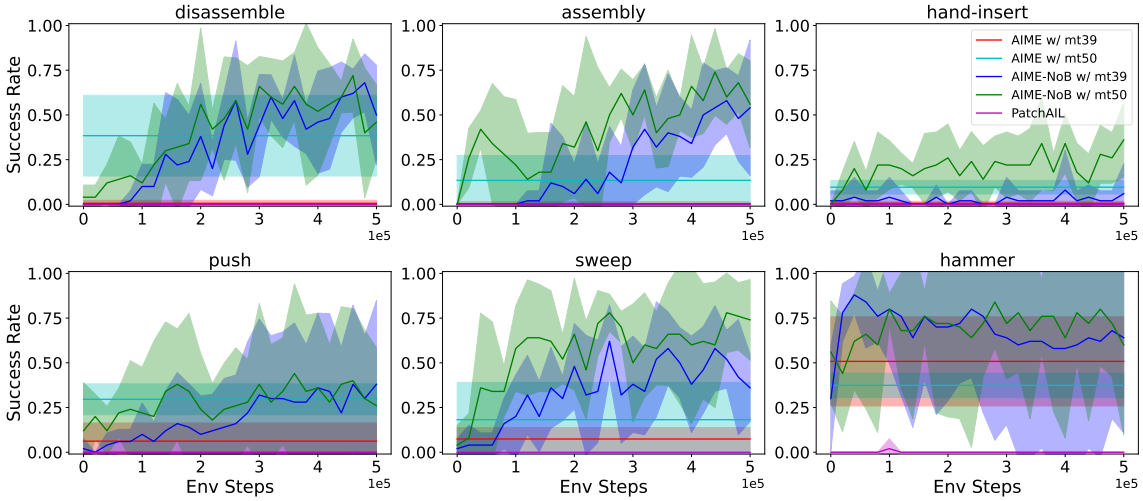


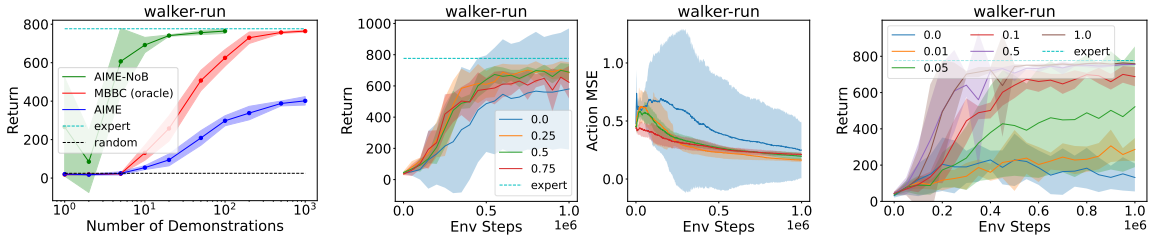
Figure 3: Benchmark results and 6 Meta-World tasks. Trajectories are only counted as success when it success at the last time steps and the success rates are calculated with 10 policy rollouts. The results are averaged across 5 seeds with the shade region representing 95% CI.

For the VIPER model, in the original paper, the authors first pretrain a VQ-GAN (Esser et al., 2021) and then train a GPT-style auto-regressive model in the quantised space for prediction. For simplicity of the implementation, in this paper, we consider training an unconditioned latent world model as in Seo et al. (2022b) to model the VIPER reward. We use the same RSSM architecture of the model learning for DMC, only removing the condition of the actions, and we train the VIPER model for each task separately. Especially during training, we find training such a powerful model from scratch on a small dataset can easily result in over-fitting. Thus, we empirically choose to train the model only for 500 gradient steps for DMC models and 1000 gradient steps for Meta-World models. We show evidence of overfitting in Appendix I. Due to the large scale of the ELBO, we also apply symlog (Hafner et al., 2023) when computing the VIPER reward. Another difference with the original VIPER paper is that we do not use intrinsic motivation as the exploration bonus, since the AIME loss for policy learning already provides task-related guidance for exploration. We only apply an entropy regulariser to the policy as is common practice. We further show the synergy between AIME and VIPER in Appendix J.

4.3 Benchmark Results

The benchmark results of DMC are shown in Figure 2. We compare AIME-NoB with AIME (Zhang et al., 2023), BCO(α) (Torabi et al., 2018) and PatchAIL (Liu et al., 2022a), a Generative Adversarial Imitation Learning (GAIL) style algorithm. AIME-NoB significantly outperforms the PatchAIL baseline in 7 out of 9 tasks in terms of sample efficiency. Benefiting from the pretrained world model, AIME-NoB typically can reach expert performance within 200k environment steps. Compared with BCO(α), updating the model is regularised and is benefiting more from the online interaction to resolve the EKB. Compared with AIME, AIME-NoB reliably improves performance, especially in hard tasks such as walker-run and hopper where offline AIME did not manage to make any progress.

However, AIME-NoB shows little progress on two tasks: cartpole-swingup and quadruped-run. For cartpole-swingup, we observe that the policy learns to move the cart out of the scene so that the static image yields a high likelihood from the video prediction model. A similar phenomenon was also discussed in the original VIPER paper (Escontrela et al., 2023). For quadruped-run, we conjecture that it is due to visual difficulties of a reconstruction-based model. When the quadruped is initialised on the ground, due to the symmetric structure of the robot, it is impossible to figure out which action



(a) Demonstration efficiency. (b) Ablation of α. (c) Ablation of β.

Figure 4: (a) Performance of AIME-NoB, MBBC, AIME w.r.t. different number of demonstrations. For AIME-NoB, we do not show the result for more than 100 demonstrations since it is already saturated to the expert. (b) Ablation for the choice of the regulariser ratio α. The left figure shows the mean return over 10 trajectories while the right figure shows the MSE between the inferred actions and the true actions. (c) Ablation for the choice of the weight of the value gradient loss β. All results are averaged across 5 seeds with the shaded region representing a 95% CI.

corresponds to which leg, and it easily leads the action inference process to diverge. We additionally show AIME-NoB can work on these tasks with the help of the true reward in Appendix J.

The benchmark results of Meta-World are shown in Figure 3. We choose four hard or very hard tasks (disassemble, assembly, hand-insert and push) and two medium difficult tasks (sweep and hammer). While PatchAIL fails on these tasks, AIME and AIME-NoB make progress. AIME-NoB using either of the pretrained models outperforms AIME in all tasks. For hard tasks, AIME with the mt50 model performs better than with the mt39 model due to the mt39 model's large EKB with unseen objects. But in the online setting of AIME-NoB, the two models are mostly on par. Moreover, using the mt50 models is better than using the mt39 models on average, which may imply covering diverse behaviour is more valuable than knowing the expert directly.

4.4 Ablation Results

We conduct our ablation studies on walker-run task from DMC.

How well does the proposed methods resolve knowledge barriers? In order to show how well AIME-NoB resolves the two knowledge barriers, we the same experiment as in Figure 1 by providing the agent with different numbers of demonstrations. The result is shown in Figure 4a. As we discussed before, MBBC as an oracle method that circumvents the EKB is a strict upper bound for AIME. And AIME-NoB which addresses both the EKB and DKB achieves much better results and is an upper bound for MBBC. From AIME-NoB can achieve near-expert performance with as few as 5 demonstrations for this challenging task.

Influence of the data regulariser ratio α. We set the regulariser ratio α from [0.0, 0.25, 0.5, 0.75] and plot the results in Figure 4b. The results show that enabling the regularizer ($\alpha > 0$) provides reliable improvements during training. But if we disable the regulariser by setting $\alpha = 0$, the learning exhibits high variance. In some cases, it fails to work entirely, while in others, learning only begins once sufficient new data accumulates in the replay buffer. As we discussed in Section 3.1, without the regulariser, in the early stage of the training, the model can easily overfit to the replay buffer, and it explains the early flattening phase of the training. As the training progresses, more and more data is available from the replay buffer, and it can establish the regulariser on its own, which explains the dramatic growing phase of the curves. We also plot the MSE between the inferred actions and the true actions during the training process. From that we can see that a higher regulariser ratio offers more stable inference of the actions in the early phase of training.

Influence of the value gradient loss weight β. We set the weight β from [0.0, 0.01, 0.05, 0.1, 0.5, 1.0] and plot the results in Figure 4c. The results show that a small β slows learning progress toward convergence, while a larger β improves sample efficiency without causing instability. For

the sample efficiency, since we only have 10 demonstrations, DKB dominates over EKB as shown in Figure 4a. Thus, having a larger β will make the learning much faster. Regarding stability, as discussed in 3.2, AIME loss and value gradient loss operate on different regions of the environment states, making their influence on the policy independent of each other.

5 Conclusion

In this paper, we identify two knowledge barriers, namely the EKB and the DKB, which as we show limit the performance of state-of-the-art ILfO methods using pretrained models. We thoroughly analyse the underlying cause of each barrier and propose practical solutions. Specifically, we propose to use online interaction with a data-driven regulariser to solve the EKB and surrogate reward labelling to reduce the DKB. Combining these solutions, we propose AIME-NoB and showcase its efficiency compared to state-of-the-art ILfO methods. Our ablation studies show how each knowledge barrier is addressed by the proposed solution and how their hyper-parameters influence the performance. We hope our work can shed some light on the future development of ILfO method and bring more attention to the great potential of pretrained world models.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, August 2022. URL http://arxiv.org/abs/2204.01691.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223. PMLR, July 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos, June 2022. URL http://arxiv.org/abs/2206.11795.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient Online Reinforcement Learning with Offline Data. In *ICML*, May 2023. URL http://arxiv.org/abs/2302.02948.

Philip Becker-Ehmck, Jan Peters, and Patrick van der Smagt. Switching Linear Dynamics for Variational Bayes Filtering. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 553–562. PMLR, May 2019. URL https://proceedings.mlr.press/v97/becker-ehmck19a.html.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj

Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, August 2021. URL http://arxiv.org/abs/2108.07258.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale, 2022. URL https://arxiv.org/abs/2212.06817.

Yongchao Chen, Rujul Gandhi, Yang Zhang, and Chuchu Fan. NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15880–15903, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.985. URL https://aclanthology.org/2023.emnlp-main.985.

Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers, March 2024. URL http://arxiv.org/abs/2306.06531.

Branton DeMoss, Paul Duckworth, Nick Hawes, and Ingmar Posner. DITTO: Offline Imitation Learning with World Models, February 2023. URL http://arxiv.org/abs/2302.03086.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805.

Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards A Unified Agent with Foundation Models, July 2023. URL http://arxiv.org/abs/2307.09668.

Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video Prediction Models as Rewards for Reinforcement Learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=HWNl9PAYIP.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.01268. URL https://ieeexplore.ieee.org/document/9578911/.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for Deep Data-Driven Reinforcement Learning, February 2021. URL http://arxiv.org/abs/2004.07219.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR 2020*, September 2019a. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2555–2565. PMLR, May 2019b. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*, September 2020. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models, January 2023. URL http://arxiv.org/abs/2301.04104.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *ICLR 2024*, October 2023a. URL https://openreview.net/forum?id=Oxh5CstDJU.

Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline, June 2023b. URL http://arxiv.org/abs/2212.05749.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01553. URL https://ieeexplore.ieee.org/document/9879206/.

Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models, July 2022. URL http://arxiv.org/abs/2207.05608.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *7th Annual Conference on Robot Learning*, August 2023. URL https://openreview.net/forum?id=9_8LF30mOC.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyTqHL5xg.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference for Learning Representations*, San Diego, January 2017. doi: 10.48550/arXiv.1412.6980. URL http://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes, December 2022. URL http://arxiv.org/abs/1312.6114.

Alexej Klushyn, Richard Kurle, Maximilian Soelch, Botond Cseke, and Patrick van der Smagt. Latent Matters: Learning Deep State-Space Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10234–10245. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/54b2b21af94108d83c2a909d5b0a6a50-Abstract.html.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-Online Reinforcement Learning via Balanced Replay and Pessimistic Q-Ensemble. In *Proceedings of the 5th Conference on Robot Learning*, pp. 1702–1712. PMLR, January 2022. URL https://proceedings.mlr.press/v164/lee22d.html.

Anqi Li, Byron Boots, and Ching-An Cheng. MAHALO: Unifying Offline Reinforcement Learning and Imitation Learning from Observations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, August 2023. URL http://arxiv.org/abs/2303.17156.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control, February 2023. URL http://arxiv.org/abs/2209.07753.

Minghuan Liu, Tairan He, Weinan Zhang, Shuicheng Yan, and Zhongwen Xu. Visual Imitation Learning with Patch Rewards. In *The Eleventh International Conference on Learning Representations*, September 2022a. URL https://openreview.net/forum?id=OnM3R47KIiU.

Yu-Ren Liu, Yi-Qi Hu, Hong Qian, Chao Qian, and Yang Yu. ZOOpt: Toolbox for Derivative-Free Optimization. *Science China Information Sciences*, 65(10):207101, s11432–021–3416–y, October 2022b. ISSN 1674-733X, 1869-1919. doi: 10.1007/s11432-021-3416-y. URL http://arxiv.org/abs/1801.00329.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=IEduRUO55F.

Parsa Mahmoudieh, Deepak Pathak, and Trevor Darrell. Zero-Shot Reward Specification via Grounded Natural Language. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 14743–14752. PMLR, June 2022. URL https://proceedings.mlr.press/v162/mahmoudieh22a.html.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, March 2023. URL https://openreview.net/forum?id=NJtSbIWmt2T.

Mitsuhiko Nakamoto, Yuexiang Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=GcEIvidYSw.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, March 2022. URL http://arxiv.org/abs/2203.02155.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The Unsurprising Effectiveness of Pre-Trained Vision Models for Control, March 2022. URL http://arxiv.org/abs/2203.03580.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. pp. 24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sai Rajeswar, Pietro Mazzaglia, Tim Verbelen, Alexandre Piché, Bart Dhoedt, Aaron Courville, and Alexandre Lacoste. Mastering the Unsupervised Reinforcement Learning Benchmark from Pixels. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28598–28617. PMLR, July 2023. URL https://proceedings.mlr.press/v202/rajeswar23a.html.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A Generalist Agent, May 2022. URL http://arxiv.org/abs/2205.06175.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models, May 2014. URL http://arxiv.org/abs/1401.4082.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, June 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to Explore via Self-Supervised World Models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8583–8592. PMLR, November 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked World Models for Visual Control, June 2022a. URL http://arxiv.org/abs/2206.14244.

Younggyo Seo, Kimin Lee, Stephen James, and Pieter Abbeel. Reinforcement Learning with Action-Free Pre-Training from Videos. In *ICML*, volume 162, pp. 19561–19579. PMLR, 2022b. doi: 10.48550/arXiv.2203.13880.

Rutav Shah and Vikash Kumar. RRL: Resnet as representation for Reinforcement Learning, November 2021. URL http://arxiv.org/abs/2107.03380.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating Situated Robot Task Plans using Large Language Models, September 2022. URL http://arxiv.org/abs/2209.11302.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral Cloning from Observation, May 2018. URL http://arxiv.org/abs/1805.01954.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative Adversarial Imitation from Observation, June 2019. URL http://arxiv.org/abs/1807.06158.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023a. URL http://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023b. URL http://arxiv.org/abs/2307.09288.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. Dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, November 2020. ISSN 2665-9638. doi: 10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. ChatGPT for Robotics: Design Principles and Model Abilities, July 2023. URL http://arxiv.org/abs/2306.17582.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Avnish Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning, June 2021. URL http://arxiv.org/abs/1910.10897.

Xingyuan Zhang, Philip Becker-Ehmck, Patrick van der Smagt, and Maximilian Karl. Action Inference by Maximising Evidence: Zero-Shot Imitation from Observation with World Models. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=WjlCQxpuxU.

A Related Work

Imitation Learning from Observation. ILfO (Torabi et al., 2018; 2019; DeMoss et al., 2023; Li et al., 2023; Baker et al., 2022; Zhang et al., 2023; Liu et al., 2022a) has become more popular in recent years due to their potential to utilise internet-scale videos for behaviour learning. Most of the previous works (Torabi et al., 2018; 2019; Li et al., 2023) study the problem only with the true state as observation. Recent works (DeMoss et al., 2023; Baker et al., 2022; Zhang et al., 2023; Liu et al., 2022a) have started to shift toward image observations as a more general setting. Our work is a continuation of this journey.

Pretrained Models for Decision-Making. Inspired by the tremendous progress made in recent years in CV and NLP fields with the power of pretrained models, the decision-making community is also trying to follow the trend. Most recent works focus on the use of Large Language Model (LLM) for decision-making. A prompted model is used for producing trajectories and plans (Chen et al., 2024; Huang et al., 2022; Ahn et al., 2022; Di Palo et al., 2023), code (Vemprala et al., 2023; Liang et al., 2023; Singh et al., 2022; Chen et al., 2023; Huang et al., 2023) or for modifying the reward (Ma et al., 2023; Mahmoudieh et al., 2022). There are also other people studying the benefit of pretrained visual models for visuomotor tasks (Shah & Kumar, 2021; Majumdar et al., 2023; Hansen et al., 2023b; Parisi et al., 2022) while others try to train large policy networks directly with transformers (Vaswani et al., 2017) and huge datasets (Brohan et al., 2022; Brohan et al.; Reed et al., 2022). However, there is only little attention being put on pretrained world models (Zhang et al., 2023; Rajeswar et al., 2023; Sekar et al., 2020), which are natively developed by the model-based decision-making community and perfectly fit into the pretraining and finetuning paradigm. Our work explores this overlooked domain and showcases its potential.

B Limitations

First, the data-driven regulariser is not practical when the model is pretrained on huge datasets – cf. foundation models popular in the fields of CV and NLP. Reducing the amount of data needed for the regulariser could greatly improve the usability of the method. Second, although having pretrained models is beneficial, having too many pretrained components can be detrimental for model selection. Especially in AIME-NoB, the world model and the VIPER model share a very similar interface. Designing a shared model that can serve both interfaces could ease the use of the method. Last but not least, due to the high demand of computing resources, we only study the pretrained world model on a very small scale. It will be an interesting direction to study these model at larger scales.

C Details for the value function update

The TD(λ) estimation $V_\xi^\lambda(s)$ is computed by a weighted sum of the multi-step TD targets, i.e.

$$V_\xi^\lambda(s_t) = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}V_\xi^{(n)}(s_t) \tag{6}$$
$$\text{with } V_\xi^{(n)}(s_t) = \sum_{t'=t+1}^{t+n}\gamma^{t'-t-1}r_{t'}^{\text{VIPER}} + \gamma^n V_\xi(s_{t+n}).$$

Using this estimate, we optimise our value function by minimising the MSE, i.e.

$$\xi^* = \operatorname*{argmin}_\xi (V_\xi(s_t) - V_{\xi'}^\lambda(s_t))^2. \tag{7}$$

As is common practice, we use a target value network with parameters ξ' to stabilise training, whose parameters are updated using Polyak averaging with a learning rate τ in every iteration.

Algorithm 1 AIME-NoB

Input: Embodiment dataset D_{body}, Demonstration dataset D_{demo}, Pretrained world model parameters ϕ, θ, Pretrained VIPER model p_ν, regulariser ratio α, value gradient weight β, batch size B
Initialise policy and critic parameters ψ, ξ randomly.
for $i = 1$ **to** policy pretraining iterations **do**
  Draw a batch of demonstrations $o_{1:T} \sim D_{\text{demo}}$
  Update policy parameters ψ with Equation (3).
end for
Initialize $D_{\text{online}} \to \emptyset$.
for $i = 1$ **to** Environment Interaction budget **do**
  Collect a new episode $\{o_{1:T}, a_{1:T}\}$ with the current policy π_ψ
  Estimate reward using VIPER $r^{\text{VIPER}}_{1:T} = p_\nu(o_{1:T})$
  Append $\{o_{1:T}, a_{1:T}, r^{\text{VIPER}}_{1:T}\}$ to D_{online}
  # Update world model
  Draw $\alpha \cdot \text{B}$ samples $b_{\text{body}} \sim D_{\text{body}}$
  Draw $(1-\alpha) \cdot \text{B}$ samples $b_{\text{online}} \sim D_{\text{online}}$
  Define combined batch $b = b_{\text{body}} \cup b_{\text{online}}$
  Finetune model with batch b using Equation (4).
  # Update policy
  Sample a batch from D_{demo}
  Update policy parameters ψ with Equation (5).
  Update value function parameters ξ with Equation (7).
end for

D Compute Resources

All the experiments are run on a local cluster with a few A100 and RTX8000 instances. All the experiments are tuned to use less than 10GB of GPU memory so that they can run in A100 MIG. World models pretraining requires about 24 GPU hours, while VIPER models require negligible time for training. Each DMC experiment requires about 40 GPU hours while each Meta-World experiment requires about 24 GPU hours.

E Algortihm

F Hyper-parameters

Here, we document the detailed hyper-parameters for all the trained models in Table 1.

G Source of Datasets

We use the expert trajectories from Liu et al. (2022a) at `https://osf.io/4w69f/?view_only=e29b9dc9ea474d038d533c2245754f0c`. The authors didn't provide a License for their dataset. Besides, we use the replay buffer dataset from Hansen et al. (2023a) at `https://huggingface.co/datasets/nicklashansen/tdmpc2/tree/main/mt80`. The authors provide the dataset under the MIT License. Moreover, we use the replay buffer dataset from Zhang et al. (2023) at `https://github.com/argmax-ai/aime/tree/main/datasets`. The authors provide the dataset under the CC BY 4.0 License.

Table 1: AIME-NoB hyper-parameters use for each benchmark.

	DMC	Meta-World
World Model		
CNN structure	Ha & Schmidhuber (2018)	Hafner et al. (2023)
CNN width	32	48
MLP hidden size	512	640
MLP hidden layer	2	3
MLP activations	LayerNorm + Swish	
Deterministic latent size	512	1024
Stochastic latent size	30	
Free nats	1.0	
KL balancing	0.8	
Policy		
Hidden size	128	
Hidden layer	2	
Activation	ELU	
Distribution	Tanh-Gaussian	
Value network		
Hidden size	128	
Hidden layer	2	
Activation	ELU	
Target EMA decay	0.95	
Training		
Batch Size	50	16
Horizon	50	64
Total Env steps	1M	500k
Model Pretraining Gradient steps	200k	
Update ratio	0.1	
Gradient clip	100	
Policy entropy regulariser weight	1e-4	
Model learning rate	3e-4	
Policy learning rate	3e-4	
Value network learning rate	8e-5	
Discount factor γ	0.99	
TD-Lambda parameter λ	0.95	
Imagine horizon	15	
AIME-NoB specific		
Policy pretraining iterations	2000	
Data-driven regulariser ratio α	0.5	
Value gradient loss weight β	0.1	

H Details for Resetting Meta-World Tasks

To generate the image observation datasets from the TD-MPC2 replay buffer (Hansen et al., 2023a), we modify the Meta-World codebase to reset the environment to the initial state of the trajectory from the first observation. Luckily, the starting position of the robot arm is always the same for each task, so that we do not need to apply inverse kinematics to solve for the initial pose of the robot arm. For the object and the target position, for most of the tasks, the internal reset position can be computed by making a constant shift on the object position and the target position in the observations. There are, however, also a few edge cases which we handle differently.

In button-press-topdown and button-press-topdown-wall, the object's true position only appears in the observation upon the second time step, presumably due to some simulator delay in the resetting process. So for these two tasks, the initial state is reset by the second observation.

For basketball and box-close, it seems like there is some internal collision detection that will alter the object and robot position after the task is reset, so computing the exact reset value from the observation is not possible. For these two tasks, we instead resort to a search-based method. To be specific, we use a gradient-free optimiser from (Liu et al., 2022b) to search over the resetting space of the object and find the reset position that minimises the L2 distance with the true observation.

More details of the implementation can be found in the code.

I Overfitting of the VIPER Models

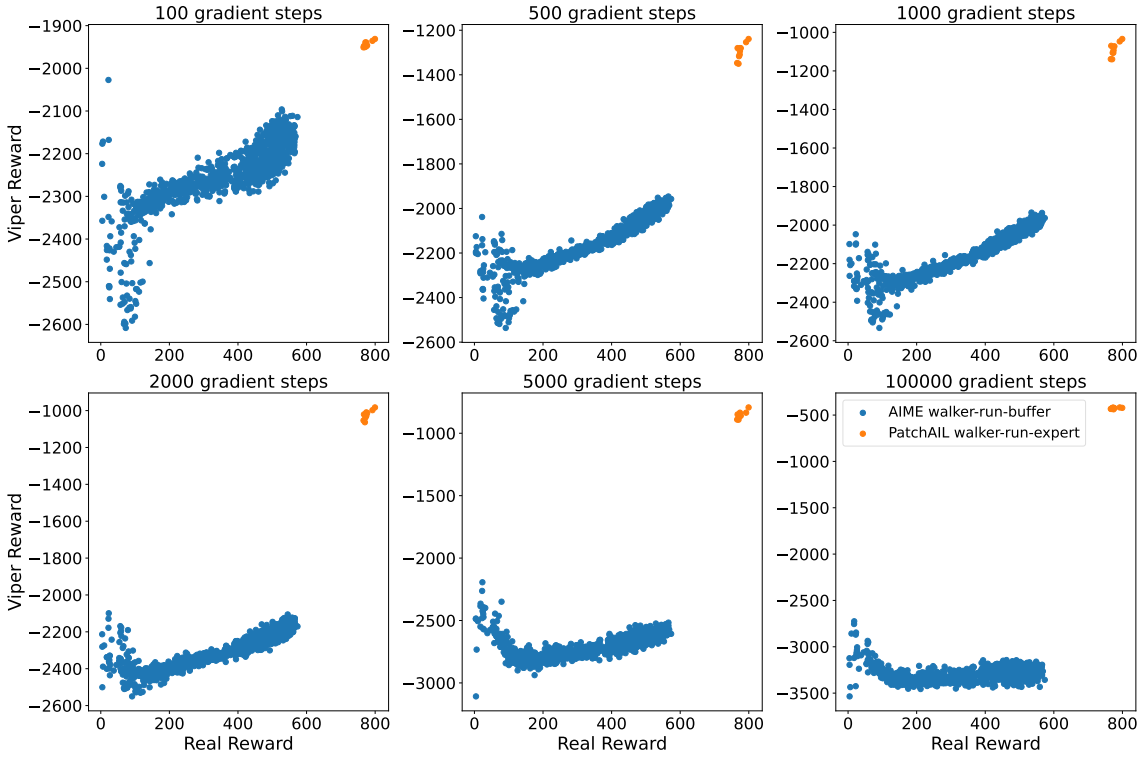


Figure 5: Correlation of the VIPER reward and the real reward with models trained with different numbers of gradient steps. Each point represents one trajectory. We can clearly see the model gradually overfitting and losing the correlation with the real reward when training for more than 1000 gradient steps.

To better illustrate the overfitting problem for VIPER models and justify our choice of training fewer iterations, we train the VIPER models for a varying number of gradient steps and evaluate the correlations between the VIPER reward and the true reward on both the expert dataset from PatchAIL, where the VIPER model is trained on, and the replay buffer dataset from Zhang et al. (2023).

Specifically, we train the same VIPER model with {100, 500, 1000, 2000, 5000, 100000} gradient steps and plot the result in Figure 5. As we can clearly see, when training with less than or equal to 1000 gradient steps, VIPER reward has a very nice correlation with the true reward, with the middle-range performance even like a linear correlation. The best model could be selected from

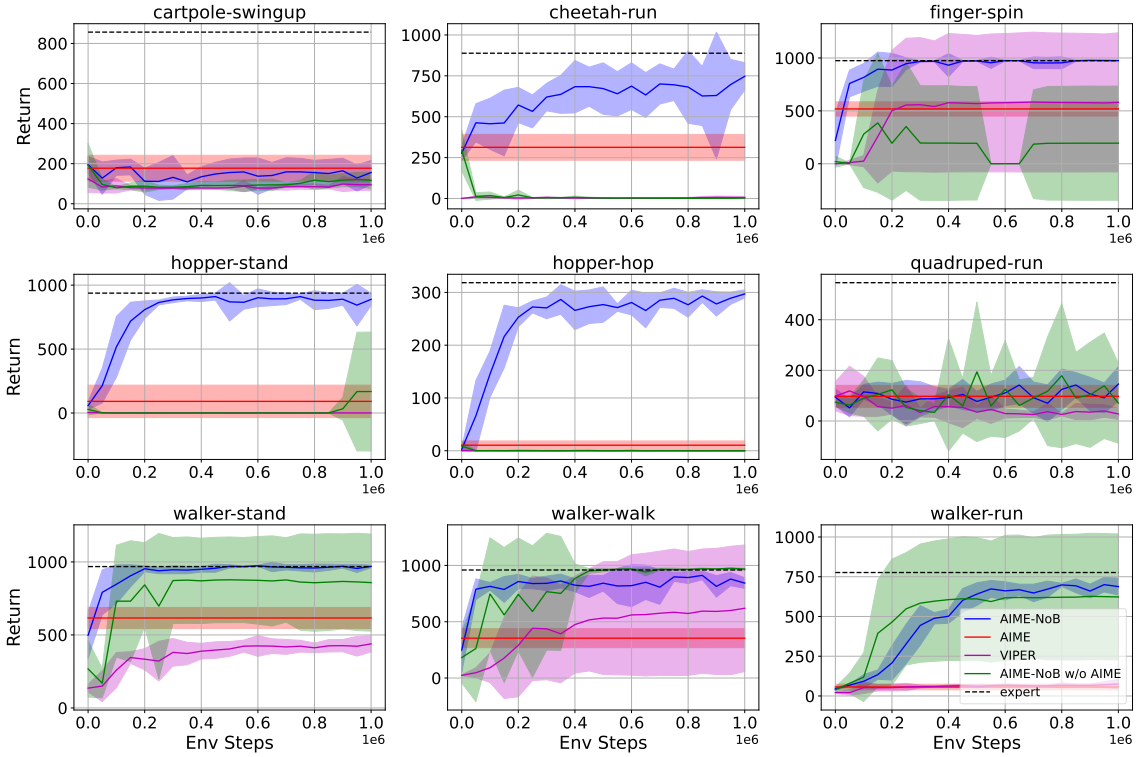


Figure 6: Additional ablation on DMC tasks by exploring the synergy between AIME and VIPER model. Return are calculated by running the policy 10 times with the environment and taking the average return. The results are averaged across 5 seeds with the shade region representing 95% CI.

500 and 1000 gradient steps. However, as we train the model for longer, the VIPER reward for the expert trajectories is boosted even higher, and as a side effect, it also relatively boosts up the VIPER reward for low-performance trajectories. This is because, when overfitting the expert trajectories, the model increases the marginal likelihood of all the observations in the expert trajectories to a higher value, which also includes a few frames of the robot lying on the ground at the very beginning of each trajectory after reset. For these low-performance trajectories, the robot remains mainly stuck around the initial position and struggles on the ground. This artifact of the overfitted VIPER reward creates a sharp local maximum in the low-performance region that the agent can hardly get away from.

J Additional Experiments

Synergy between AIME and VIPER model. We also find there is a synergy between AIME and VIPER model. As we showed in Appendix I, one inherent problem of VIPER reward is that it not only incentivises the expert behaviour as the optimal, but also a stationary behaviour with very low reward as a local maxima. In order to work with the VIPER reward, the agent needs to have the ability to escape from the local maxima region. AIME offers the IL loss to imitate the expert demonstrations and can achieve decent performance even when pretrained offline, which helps to escape the local maxima. To better show the synergy, we provide additional ablation results with the VIPER reward in Figure 6. In the experiments, we include two other variants: the AIME-NoB w/o AIME is to remove the AIME IL loss from the online policy learning, so that the policy is pretrained by AIME loss but finetuned with only RL loss from the VIPER reward; while the VIPER is following the implementation in the original VIPER paper with RL loss on both the VIPER reward for the task and intrinsic reward for exploration. From the result, we can clearly see that without the help

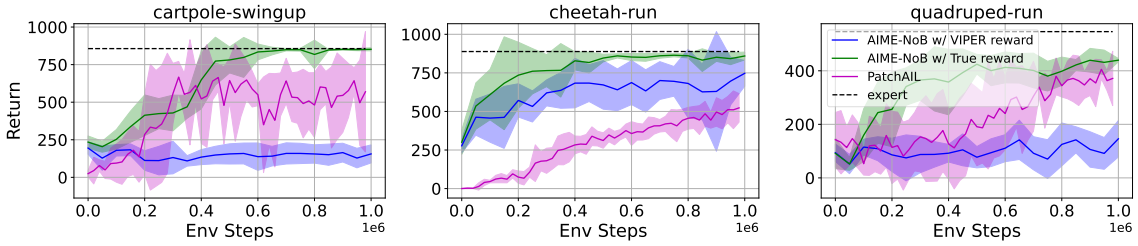


Figure 7: Additional benchmark results on 3 DMC tasks with an additional variant of AIME-NoB with the true reward. Return are calculated by running the policy 10 times with the environment and taking the average return. The results are averaged across 5 seeds with the shade region representing 95% CI.

of AIME IL loss, VIPER reward cannot reliably motivate the agent to learn good behaviours. Even when in walker tasks, the w/o AIME variant can solve the tasks to a certain extent; it depends strongly on the random seeds. In conclusion, the good performance of AIME-NoB cannot be achieved by either AIME IL loss or VIPER RL loss alone but by a combination of both.

Improving AIME-NoB with better rewards. We show additional results on the 3 not-so-well-performing DMC tasks, namely cartpole-swingup, cheetah-run and quadruped-run, in 7. In the plot we add a new variant using the true reward from the environment to replace the VIPER reward. As the results show, if we had a better estimation of the surrogate reward, AIME-NoB could also achieve good performance on these tasks.