SeMAIL: Eliminating Distractors in Visual Imitation via Separated Models

Shenghua Wan [1] Yucen Wang [1] Minghao Shao [1] Ruying Chen [1] De-Chuan Zhan [1]

Abstract

Model-based imitation learning (MBIL) is a popular reinforcement learning method that improves sample efficiency on high-dimension input sources, such as images and videos. Following the convention of MBIL research, existing algorithms are highly deceptive by task-irrelevant information, especially moving distractors in videos. To tackle this problem, we propose a new algorithm - named Separated Model-based Adversarial Imitation Learning (SeMAIL) - decoupling the environment dynamics into two parts by task-relevant dependency, which is determined by agent actions, and training separately. In this way, the agent can imagine its trajectories and imitate the expert behavior efficiently in task-relevant state space. Our method achieves near-expert performance on various visual control tasks with complex observations and the more challenging tasks with different backgrounds from expert observations.

1. Introduction

Reinforcement learning enables agents to autonomously learn specific behaviors and acquire diverse skills by interacting in environments. However, the rewards used for agent training are hard to define due to the expensive professional knowledge in domains (Hadfield-Menell et al., 2017; Everitt et al., 2021). One promising direction is imitation learning (Ross et al., 2011; Osa et al., 2018), where agents learn desired behaviors directly from expert demonstrations, especially from more readily available data such as images and videos (Stadie et al., 2017; Young et al., 2020). With model-based methods introduced in this field (Baram et al., 2016; Fu et al., 2021; Rafailov et al., 2021; Hu et al., 2022), agents can sample data and train policy in the learned model, which significantly improves the sample efficiency and representation learning ability in visual observations.

[1] National Key Laboratory for Novel Software Technology , Nanjing University, Nanjing, China. Correspondence to: De-Chuan Zhan <zhandc@nju.edu.cn>.
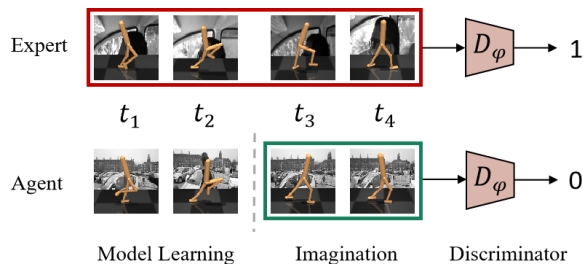
Figure 1. Illustration of convention model-based imitation learning which trains the policy in a single environment model. Real-world cases may contain a large number of distractors, especially moving distractors like walkers walking on the roadside. The model trained in these cases will tell the agent not only the result of its action but also something else irrelevant to the task. The discriminator may mistakenly use those imagined distractors to distinguish the agent and expert, even if they have similar task-relevant behaviors.

In model-based visual imitation learning algorithms, there is a common issue caused by distractors that majorly hinder learning ability in complex real-world scenarios, as shown in Figure 1. In this case, the agent already obtained task-completion information from the expert by exploring specific actions. However, the environment model contains various task-irrelevant details and will take that irrelevant information into the next predicted state. Because of the difference in task-irrelevant parts, the agent only can receive a low reward from the discriminator even if the action in the task-relevant part nicely dovetails with the target.

Recent studies attempted to exclude distractors by bottlenecking the information between latent states and observations (Peng et al., 2019) or extracting features unrelated to the task from the initial frames explicitly (Zolna et al., 2020). Other works match the policy and expert by minimizing mutual information using randomly pre-collected data in both domains (Stadie et al., 2017; Cetin & Çeliktutan, 2021). In real-world practice, these methods still have shortcomings : (1) Only address IL problems on low-dimensional inputs or images with static distractors. (2) Pre-collected data is required for aligning the agent and the expert. To close the gap, we focus on excluding task-irrelevant information without additional requirements, which benefits solving real-world cases.

When the agent executes the same action based on the same states, the task-relevant parts will have a similar transition; contrariwise, the task-irrelevant parts will have independent transitions. We claim that only the task-relevant states in observation will change consistently when the agent and expert make a specific action. Based on this finding, we make the Action-conditioned Transition (AcT) assumption. We propose a practical method named Separated Model-based Adversarial Imitation Learning (SeMAIL) based on the assumption. For model learning, we design two models to estimate the forward dynamics for these two parts and dissociate the irrelevant one from the action input. The two separated models are trained on data from both expert and agent buffer. After that, We enforce task-relevant and irrelevant latent states together to reconstruct the original observation to learn the corresponding representations. We train the agent by adversarial imitation learning for policy learning, but the policy data used in discriminator and policy training is collected only in the task-relevant latent state space.

To begin with, we evaluate our method on modified Deep-Mind Control Suite tasks (Tassa et al., 2018), which have complex real-scene video distractors in the background, and a 2-D classical control task from OpenAI Gym (Brockman et al., 2016). The results show that SeMAIL achieves near-expert performance on various tasks. Further, SeMAIL and its variants outperform compared methods on a series of tasks with different distractors from expert demonstrations. From those sides, our proposed method can capture the most helpful information (e.g. agent joints or target body) for completing tasks, especially for tasks with complex and time-correlated distractors.

To summarize our contributions in this work: (i) We propose a new approach, Separated Model-based Adversarial Imitation Learning (SeMAIL), which aims to solve IL tasks from visual observations with complex or time-correlated distractors. (ii) We provide a theoretical analysis that shows the performance gap between agent and expert in observation space can also be bounded in the task-relevant latent state space. (iii) We show the superior performance of policies learned by SeMAIL across multiple visual control tasks with complex distractors, including learning from expert demonstrations with backgrounds different from observations.

2. Related Work

Reward Estimation in Visual Imitation Learning. A practical way for imitation learning is to get the expert's supervised signal directly from high-dimensional inputs like images or videos (Finn et al., 2017; Pathak et al., 2018; Young et al., 2020; Haldar et al., 2022). Matching the agent and expert distribution based on extracting representations

from visual observations and estimating the correct reward is an inherently challenging task. Previous research generates rewards for agent training based on aligned expert and agent trajectories through proprioceptive information (Torabi et al., 2019). P-SIL (Cohen et al., 2021) uses the Sinkhorn distance (Cuturi, 2013) to estimate the reward to guide the agent learning. Recent work (Liu et al., 2022) proposes a new approach that recovers the reward based on the similarity of the image patch between the agent and expert observations, but it can not estimate the reward precisely if the observation contains much task-irrelevant information. Our method aims to provide task-relevant-based rewards so that it can solve visual imitation learning tasks with complex and time-correlated distractors.

Model-based Imitation Learning. Several works have introduced model-based approaches (Sutton & Barto, 1990; Deisenroth & Rasmussen, 2011) to imitation learning to address the problem of sampling efficiency. MAIL (Baram et al., 2016) trains the model and the discriminator on off-policy data for adversarial imitation learning, but it only focuses on solving low-dimensional state tasks. GCL (Finn et al., 2016) equips inverse reinforcement learning with a local linear dynamics model to learn a good cost function but can not directly learn from natural images. For visual inputs, V-MAIL (Rafailov et al., 2021) is a representative model-based imitation learning method that learns a surrogate model for underlying MDPs and provides a theoretical analysis of the performance bound between the agent and the expert on partially observed MDPs. MILE (Hu et al., 2022) trains a recurrent state-space model on an offline set of expert data without any interaction with the environment and can handle the image observation with a large size. All of these methods learn an environment model to approximate the underlying dynamics to improve sample efficiency in imitation learning. However, they ignore the possible distractors with their own transition dynamics, common in real-world cases, and can not prioritize task-relevant features in the learning process. If these distractors are absorbed into latent states, it will not only influence the next stage in the model learning process but further make the agent and discriminator focus on some task-irrelevant features.

Reinforcement Learning with Noisy Observations. Many recent RL studies have explored ways for better performance in environments with noise or distractors. These methods can be divided into two categories: reconstruction-free and reconstruction-based. As reconstruction-free methods, DBC (Zhang et al., 2021) learns a compact latent state by bisimulation metric to filter out distractors in the environment. InfoPower (Bharadhwaj et al., 2022) combines a variational empowerment term into the state-space model to capture task-relevant features at first. These works substitute the reconstruction's functionality with other designs. For reconstruction-based methods, Denoised MDPs (Wang et al.,

2022) decomposes the visual observation into four parts by action and reward, and constructs the corresponding models. The most similar method to our approach is TIA (Fu et al., 2021), which also designs two models to capture the task and distractor features. All these methods, including TIA, separate the task-related and distractor information through **rewards** based on human-designed prior knowledge. Our method has essential differences from TIA in model learning and policy learning. For model learning, we design a model to model task-relevant and irrelevant dynamics through conditioned actions instead of using ground-truth rewards as in TIA. For policy learning, we use a discriminator to train the agent by expert demonstrations in the task-relevant state space, while TIA trains it using rewards decoded by the learned model.

3. Preliminaries

A partially observed Markov decision process (POMDP) is an MDP in which agents must make decisions based on incomplete information (Kaelbling et al., 1998). POMDP can be formalized as a 7-tuple $\langle \mathcal{S}, \mathcal{A}, r, p, \mathcal{O}, \Omega, \gamma \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function, $\mathcal{O}$ is the observation space, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the state transition probability function, $\Omega : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \rightarrow [0, 1]$ is the observation probability function and $\gamma \in [0, 1)$ is the discount factor. In a POMDP, the agent infers the belief state with incomplete observations of the environment and uses this information to make decisions that maximize its cumulative reward. Since the agent in imitation learning with complex distractors has no access to the ground-truth state, we can regard it as a type of POMDP[1]. The agent can only infer the belief state from the historical observations and accept the supervised signal from demonstrations collected by the optimal expert's policy π^E in an inverse RL fashion.

Adversarial Imitation Learning (AIL) (Ho & Ermon, 2016; Fu et al., 2018) is a typical sort of inverse RL algorithm, which takes the advantage of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) to train a discriminator D_ψ to distinguish between the agent trajectory collected from the environment and the expert data from fixed demonstrations. AIL aims to learn a policy π to minimize the divergence between the expert and agent occupancy measures, meanwhile maximizing its entropy. It can be formulated as $\arg \min_\pi -H(\pi) + \psi^\star(\rho_\pi - \rho_{\pi_E})$, where ψ is a regularizer and $\psi^\star$ is convex conjugate of it. There are various choices for ψ (Ghasemipour et al., 2019), and we adopt the JS divergence used in GAIL (Ho & Ermon, 2016). The objective is

---

[1]More discussions about the POMDP assumption are in Appendix E.2.

as follows:

$$
\begin{aligned}
\max_\pi \min_{D_\psi} \ & \mathbb{E}_{(s,a) \sim \rho_{\mathcal{M}}^E} \Big[-\log D_\psi(s, a) \Big] \\
& + \mathbb{E}_{(s,a) \sim \rho_{\mathcal{M}}^\pi} \Big[-\log(1 - D_\psi(s, a)) \Big]
\end{aligned}
\tag{1}
$$

4. Approach

In this section, we first introduce the basic assumptions and overall architecture of SeMAIL (Section 4.1). SeMAIL separates the model learning into two branches: task-relevant, which captures the task information with action inputs, and task-irrelevant, which learns the environmental background and other distractors dynamics without action inputs. We derive the lower bound of the mutual information between the observations and the latent states for observation reconstruction. We encourage the two models to reconstruct the original observations cooperatively to maximize the mutual information (Section 4.2). For policy learning, we use the GAIL framework to optimize the policy based on the theoretical result of Theorem 4.1 (Section 4.3).

4.1. Learning Separated Models

We make a basic assumption, named Action-conditioned Transition (AcT), of the underlying mechanism of the environment, whose latent state z_t consists of the task-relevant part z_t^+ and irrelevant part z_t^- at time t. We assume that z_t^+ depends on the last task-relevant part z_{t-1}^+ and action a_{t-1}, and z_t^- only depends on the last task-irrelevant part z_{t-1}^-. These two components z_t^+ and z_t^- process **independently**[2]. Thus, we can decompose the forward dynamics into two independent transition functions such that

$$
p(z_t | z_{t-1}, a_{t-1}) = p(z_t^+ | z_{t-1}^+, a_{t-1}) p(z_t^- | z_{t-1}^-)
\tag{2}
$$

The overall architecture is shown in Figure 2(a). First, we use two separated encoders E_{ϕ^+} and E_{ϕ^-} to extract the low-dimensional embedding z_t^+ and z_t^- from original image observations o_t. Then, we design two forward transition models for learning the task and environment background dynamics separately: the *task model* $p_{\theta^+}(z_{t+1}^+ | z_t^+, a_t)$ which predicts the next task-relevant latent state, and the *background model* $p_{\theta^-}(z_{t+1}^- | z_t^-)$ which infers the next irrelevant latent state. We simultaneously learn two variational encoders $q_{\psi^+}(z_t^+ | o_t, z_{t-1}^+, a_{t-1})$ and $q_{\psi^-}(z_t^- | o_t, z_{t-1}^-)$ to infer the posterior estimation of task-relevant and irrelevant latent states, respectively. To obtain compact representations of these two latent states, we jointly optimize the loss function in Equation (3) to minimize the KL divergence between the prior and posterior estimation of z_t^+ and z_t^-. We train the separated models on the data both from expert

---

[2]More discussions about the practice of AcT assumption can be found in Appendix E.3.

(a) Model learning
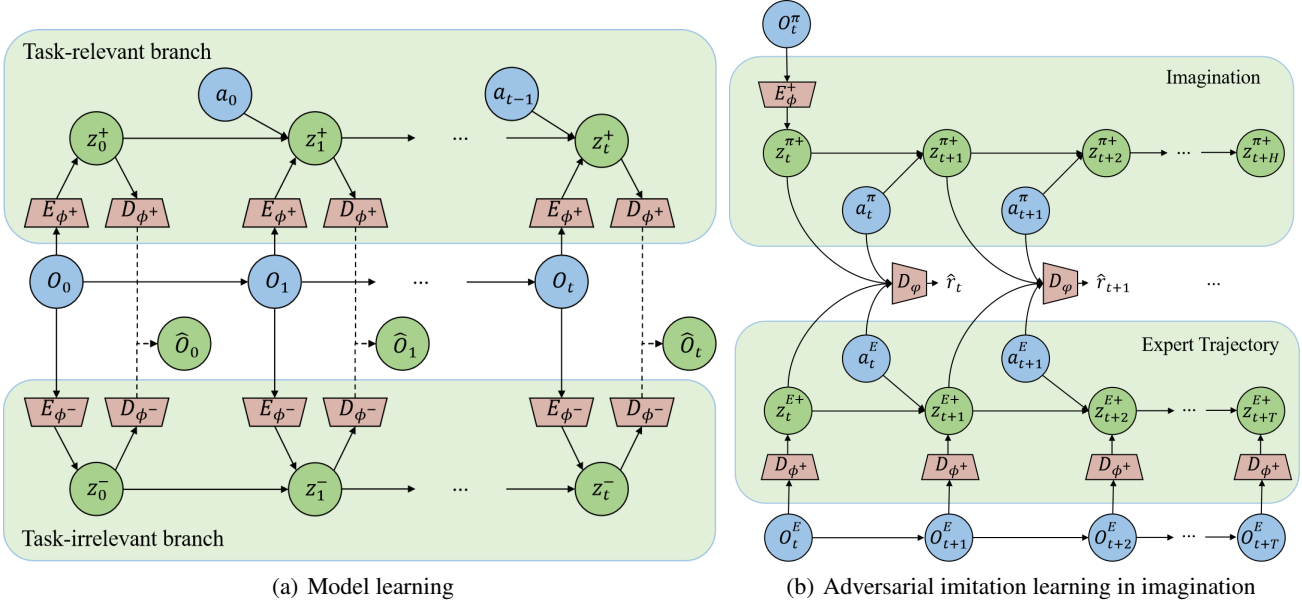
(b) Adversarial imitation learning in imagination

Figure 2. Overview of SeMAIL. (a) The model learning consists of two branches: the top branch learns $p(z_t^+|z_{t-1}^+, a_{t-1})$ for task-relevant transition, and the bottom branch learns $p(z_t^-|z_{t-1}^-)$ for distractor transition. z_t^+ and z_t^- cooperatively reconstruct the observation $\hat{o}_t$ by decoders $D_{\phi+}$ and $D_{\phi-}$. (b) For policy learning, the agent samples trajectories in the learned task-relevant latent state space and updates its policy using rewards provided by the discriminator D_ψ.

and agent buffers. The derivation of the loss function below is in Appendix B.

$$
\begin{aligned}
\mathcal{L}_{\widetilde{\mathcal{M}}} = \ & \mathbb{E}_{(o_\tau, a_\tau) \sim \mathcal{B}_\pi \cup \mathcal{B}_E} \\
& \left[\sum_{t=1}^{T} \mathbb{E}_{q(z_{t-1}^+|o_{1:t-1}, a_{1:t-2})q(z_{t-1}^-|o_{1:t-1})} \right. \\
& \left(\mathbb{D}_{\mathrm{KL}}\left[q_{\psi-}(z_t^-|o_t, z_{t-1}^-)||p_{\theta-}(z_t^-|z_{t-1}^-) \right] + \right. \\
& \left. \left. \mathbb{D}_{\mathrm{KL}}\left[q_{\psi+}(z_t^+|o_t, z_{t-1}^+, a_{t-1})||p_{\theta+}(z_t^+|z_{t-1}^+, a_{t-1}) \right] \right) \right]
\end{aligned}
\tag{3}
$$

4.2. Joint Observation Reconstruction

Most of previous model-based reinforcement learning methods consider auxiliary reconstruction-loss to optimize the observation encoder (Hafner et al., 2020; Rafailov et al., 2021). They formulate this loss as maximizing the mutual information $I(o_t; z_t)$ between the observations and latent states. We can optimize it by maximizing its BA lower bound (Barber & Agakov, 2003):

$$
I(o_t; z_t) \geq \mathbb{E}_{p(o_t, z_t)}[\ln q_\phi(o_t|z_t)] + \mathcal{H}(p(o_t))
$$

Because the observations do not depend on the latent state, optimization only considers maximizing the first term. How-

ever, enforcing the decoder to recover the whole observations from the encoded latent states will introduce bias when observations contain distractors. For imitation learning, it is catastrophic if the discriminator focuses on some task-irrelevant information and results in providing misleading reward to the agent. To avoid the latent state trained from all the details in observation, we maximize the mutual information between the observation o_t and the combination of the task-relevant latent state z_t^+ and the irrelevant latent state z_t^- as follows:

$$
I(o_t; z_t^+, z_t^-) \geq \mathbb{E}_{p(o_t, z_t^+, z_t^-)}[\ln q_\phi(o_t|z_t^+, z_t^-)]
$$

To implement it, we design two decoders $D_{\phi+}$ and $D_{\phi-}$. With the given learned representation z_t^+ and z_t^-, decoder $D_{\phi+}$ and $D_{\phi-}$ will output the task-relevant visual component $\hat{o}_t^+$ with mask M_t^+ and the irrelevant visual component $\hat{o}_t^-$ with mask M_t^-. We use a 2D-convolutional layer to generate the final image mask M_t from these two masks, a technique commonly utilized in prior works (He et al., 2017; Fu et al., 2021). These two visual components cooperatively recover the original observation at time t, formulated as $\hat{o}_t = M_t \odot \hat{o}_t^+ + (1 - M_t) \odot \hat{o}_t^-$.

Reconstructing the original image observation from two latent states may result in z_t^+ dominating the whole reconstruction process and thus containing some irrelevant information at time t. We assume that the task-relevant information is only a small proportion of the observation, which is also

posed in (Fu et al., 2021; Wang et al., 2022). To avoid the learning collapse of z_t^- and too much information captured by z_t^+, we design an additional observation decoder $q_{\tilde{\phi}}$ following TIA to decode the whole observation. This term is designated as background-only reconstruction (BoR), with the aim that the non-controllable latent state z_t^- can recover the task-irrelevant background as much as possible. The reconstruction loss[3] can be written as:

$$
\begin{aligned}
\mathcal{L}_O = \ &\mathbb{E}_{(o_\tau, a_\tau) \sim \mathcal{B}_\pi \cup \mathcal{B}_E} \\
&\left[\sum_{t=1}^{T} \left(\mathbb{E}_{\substack{q(z_t^+ | o_{1:t}, a_{1:t-1}) \\ q(z_t^- | o_{1:t})}} [\ln q_{\phi^+, \phi^-}(o_t | z_t^+, z_t^-)] \right. \right. \\
&\left. \left. + \lambda \mathbb{E}_{q(z_t^- | o_{1:t})} [\ln q_{\tilde{\phi}}(o_t | z_t^-)] \right) \right]
\end{aligned}
$$

(4)

where λ is a hyper-parameter that controls the weight of the BoR term. In contrast to TIA, the task-irrelevant hidden state z_t^- in the expectation of BoR loss does not rely on actions, as postulated by the AcT assumption.

4.3. Policy Learning

In this section, we provide a theoretical upper bound on the performance gap between the expert and the policy trained on the task-relevant transition model. The agent can perform planning and adversarial imitation learning in task-relevant latent states and reach a similar performance to the expert theoretically. Based on the Lemma 1 stated in V-MAIL (Rafailov et al., 2021), we extend the policy and model deviation from the original dynamics of MDPs to the task-relevant dynamics. Suppose there is a model $\widetilde{\mathcal{M}}^+$ which approximates the underlying task-relevant MDP $\mathcal{M}^+$ such that $\epsilon = \mathbb{D}_{TV}^{\max}(\widetilde{p}(s,a), p(s,a)), \forall(s,a)$. Then, the expected returns gap between expert π^E and learned policy π can be bounded as

$$
\begin{aligned}
|J(\pi^E, \mathcal{M}^+) - J(\pi, \widetilde{\mathcal{M}}^+)| \leq \\
\frac{R_{\max}}{1-\gamma} \mathbb{D}_{TV}(\rho_{\widetilde{\mathcal{M}}^+}^\pi, \rho_{\mathcal{M}^+}^E) + \frac{\epsilon \cdot R_{\max}}{(1-\gamma)^2}
\end{aligned}
$$

(5)

where $R_{\max} = \max_{(s,a)} \mathcal{R}(s,a), \forall(s,a)$ is the maximum of the reward in the MDP with task-relevant dynamics. We need to minimize the divergence $\mathbb{D}_{TV}(\rho_{\widetilde{\mathcal{M}}^+}^\pi, \rho_{\mathcal{M}^+}^E)$ to achieve near-expert performance. We provide a theory that this divergence can also be bounded in the task-relevant latent state space $\mathcal{Z}^+$, which makes our proposed SeMAIL method available under the adversarial imitation learning framework.

Theorem 4.1. *Consider a POMDP $\mathcal{M}$ with high-dimensional inputs such as images. Let s_t be the ground-truth state and z_t be the latent representation of the whole*

observation o_t. z_t^+ and z_t^- represent for the task-relevant and irrelevant component, respectively, which meets the condition of $p(s_t | z_t, a_t) = p(s_t | z_t^+, a_t) p(s_t | z_t^-)$ based on the AcT assumption mentioned in Section 4.1. Then, we have

$$
\begin{aligned}
\mathbb{D}_f(\rho_{\mathcal{M}}^\pi(o,a) || \rho_{\mathcal{M}}^E(o,a)) \leq &\mathbb{D}_f(\rho_{\mathcal{M}}^\pi(s,a) || \rho_{\mathcal{M}}^E(s,a)) \\
\leq &\mathbb{D}_f(\rho_{\mathcal{M}}^\pi(z^+,a) || \rho_{\mathcal{M}}^E(z^+,a))
\end{aligned}
$$

(6)

where $\mathbb{D}_f$ is a generic f-divergence.

Theorem 4.1 illustrates that the divergence of occupancy measures in the observation space can be upper-bounded in the learned task-relevant latent state space. The proof is in A.1. As the learned model $\widetilde{\mathcal{M}}^+$ well approximates the true MDP $\mathcal{M}^+$, we can apply the adversarial imitation learning method in $\mathcal{Z}^+$ space to minimize the divergence $f(\rho_{\widetilde{\mathcal{M}}^+}^\pi(z^+,a) || \rho_{\mathcal{M}^+}^E(z^+,a))$ and reduce the performance gap between the agent and the expert. The objective is

$$
\begin{aligned}
\max_\pi \min_{D_\psi} \ &\mathbb{E}_{(z^+,a) \sim \rho_{\mathcal{M}}^E} \left[-\log D_\psi(z^+,a) \right] \\
&+ \mathbb{E}_{(z^+,a) \sim \rho_{\widetilde{\mathcal{M}}}^\pi} \left[-\log(1 - D_\psi(z^+,a)) \right]
\end{aligned}
$$

(7)

The discriminator D_ψ serves as an approximated reward function that gives an estimated reward r^t of the agent's state-action pair at time t. The agent uses the pseudo reward to fit the value function bootstrapped and maximizes the expected return. To improve sample efficiency, we only train the discriminator on the fixed expert's demonstrations and the agent's rollouts in the learned model. The agent only samples actions based on the task-relevant state.

Action model: $a_t \sim \pi(a_t | z_t^+)$

Value model: $v(z_t^+) \approx \mathbb{E}_{\pi(\cdot | z_t^+)} \left[\sum_{k=t}^{T} \gamma^{k-t} \log D_\psi(z_k^+, a_k) \right]$

The policy learning process is presented in detail in Figure 2(b).

5. Experiments

In our experiments part, we aim to answer the following questions:

1. How good is SeMAIL's performance on learning tasks with complex distractors observations?

2. What are the roles played by the critical designs of SeMAIL in learning?

3. Could SeMAIL perform well with distractors that never appear in expert observations?

4. How well does the task-relevant space imagination benefit adversarial imitation learning?

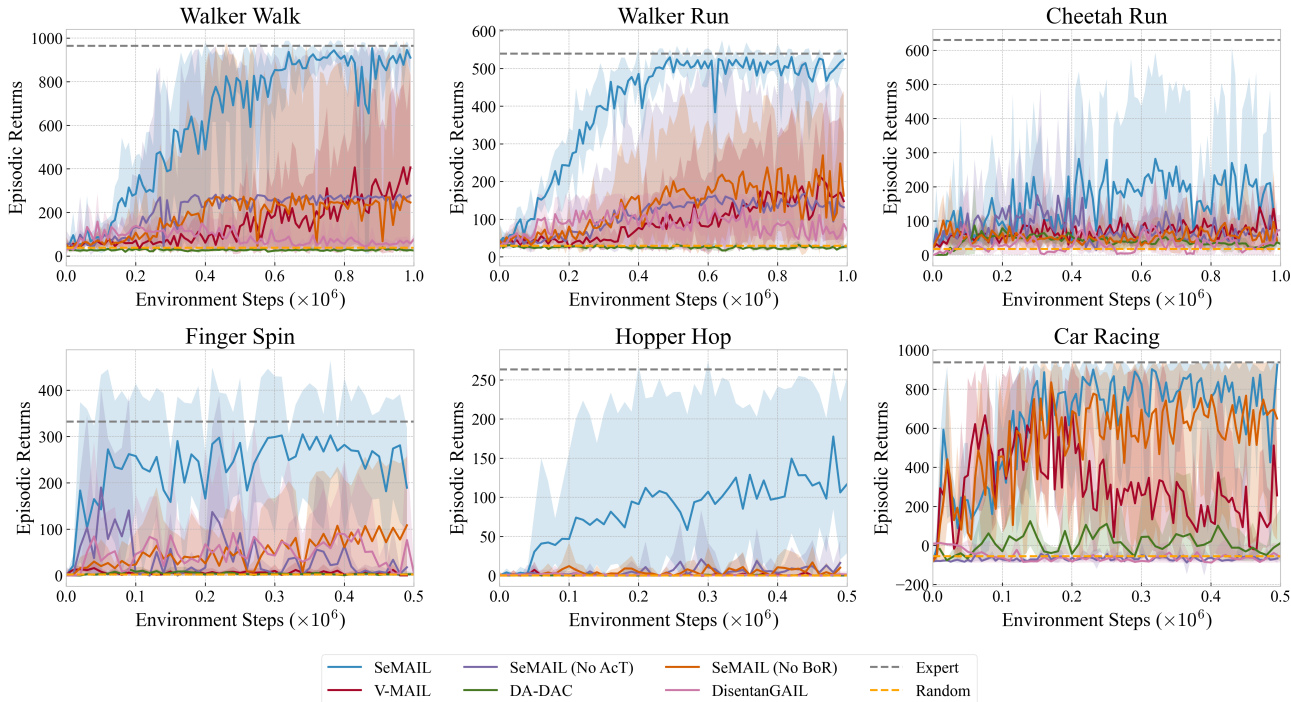[3]The derivation of the loss function is in Appendix B

Figure 3. Evaluation results of our method SeMAIL and the baselines over four seeds in six visual control tasks. The solid curves present the average episodic returns, and the shaded region represents the range of performance under different runs. SeMAIL consistently outperforms the three compared methods in almost all environments.

Environments and Expert's demonstrations. We test our algorithm on six visual control tasks, *i.e.*, five locomotion tasks from DeepMind Control (DMC) Suite (Tassa et al., 2018) with videos under the class "driving car" of the Kinetics dataset (Kay et al., 2017) as background, and another Car Racing task from OpenAI Gym (Brockman et al., 2016). These videos are grayscaled as in DBC (Zhang et al., 2021). Instead of rendering the images from low-dimensional states or adding the distractors on the pure observations, we train RL policies from complex visual inputs until optimal as experts. Then we use the expert to collect demonstrations from the above environments directly. More details about environments and demonstrations are in Appendix C.

Baselines. We design three versions of SeMAIL and compare them with baselines against three types of visual imitation learning approaches: model-based, data-augmented, and feature-space-shared.

- **SeMAIL**: The full proposed method.

- **SeMAIL (No AcT)**: SeMAIL without the action-free constraint in the task-irrelevant model. The forward dynamics and posterior encoder of the task-irrelevant model are modified as $p_{\theta^-}(z_t^- | z_{t-1}^-, a_{t-1})$ and $q_{\psi^-}(z_t^- | o_t, z_{t-1}^-, a_{t-1})$, respectively.

- **SeMAIL (No BoR)**: SeMAIL without the background-only reconstruction loss.

- **V-MAIL**: The variational model-based adversarial imitation learning method (Rafailov et al., 2021).

- **DA-DAC**: The DrQ data augmentation (Yarats et al., 2021) version of Discriminator Actor Critic (Kostrikov et al., 2019) used in (Rafailov et al., 2021).

- **DisentanGAIL**: The feature-space-shared IRL method that regularizes the latent representation with mutual information constraints (Cetin & Çeliktutan, 2021).

5.1. How good is SeMAIL's performance on learning tasks with complex distractors observations?

In Figure 3, we show the performance curves of all six visual control tasks. It is clear that SeMAIL outperforms all the baseline methods and enables high sampling efficiency in most of these tasks. SeMAIL achieves near-expert performance in Finger and Walker tasks while other methods struggle to solve them. On the Hopper Hop environment, SeMAIL has an outstanding average performance with a wider range of variance than the baselines. That is because the compared methods fail to learn the expert behavior, while for SeMAIL, there are more successful and fewer failed
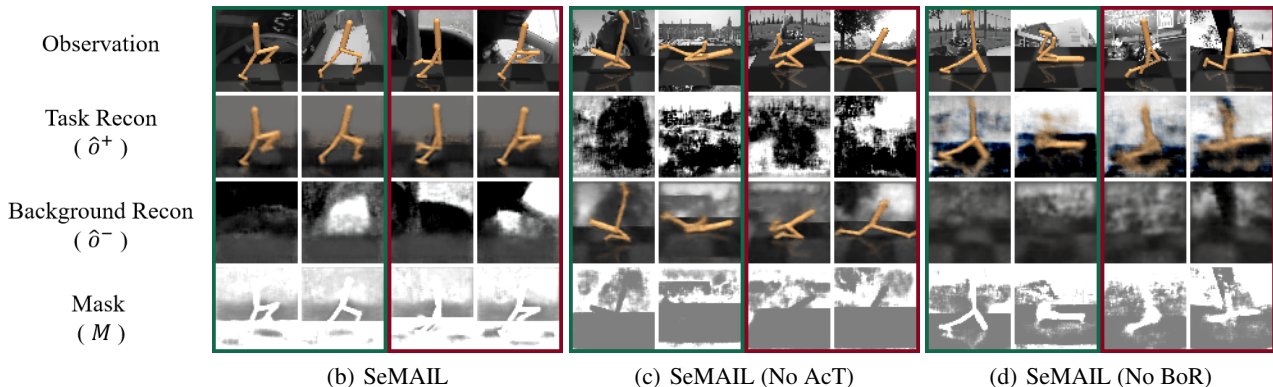
Figure 4. Reconstruction results of SeMAIL and its ablated variants for different semantics in observations from the agent (green) and the expert (red). (a) SeMAIL can successfully separate the task-relevant part and irrelevant part from observations and reconstruct them. (b) SeMAIL (No AcT) reverses the reconstructions for these two parts. There is no task information in $\hat{o}^+$. (c) SeMAIL (No BoR) can reconstruct $\hat{o}^+$ and $\hat{o^-}$ well but contain some task-irrelevant information in the former.

runs. In the Car Racing environment, V-MAIL obtains high episodic returns in the early stage, but the performance quickly drops to a value lower than SeMAIL as training goes on because the discriminator captures irrelevant parts in observations and distinguishes the agent from the expert.

The performance on all six tasks has a considerable drop if we ignore the AcT assumption. Notably, the performances of SeMAIL (No AcT) become sharply unstable in Walker Walk and Walker Run. It indicates that dissociating task-irrelevant part from actions is crucial for task completion when some distractors have underlying dynamics. Removing BoR loss also leads to a significant performance drop on several tasks, except for Car Racing, since the distractors in this environment are not visually visible. Although removing AcT or BoR component, the average performances of SeMAIL still outperform the DA-DAC and DisentanGAIL in almost all environments. The quantitative result is in Table 2.

5.2. What are the roles played by the critical designs of SeMAIL in learning?

To further study the factors that cause the performance drop on SeMAIL (No AcT) and SeMAIL (No BoR), we visualize reconstructions of different semantics of observations from the agent and the expert in Figure 4. SeMAIL learns a distinct mask and reconstructs the observations well for both tasks and backgrounds. SeMAIL (No AcT) fails to learn an accurate mask and its two parts of the reconstruction are reversed, implying that the AcT assumption is crucial for extracting the corresponding features from o^+ and o^-. For SeMAIL (No BoR), the task model captures the task-relevant information in the correct semantics but also contains some background information, which may lead to poor

performance results. Based on these ablation test results, the action-conditioned transition can help the model capture the corresponding representation with task-relevant information and majorly improves its learning ability. Background-only reconstruction tries to recover the whole background as much as possible via non-controllable latent state z^- and can filter some irrelevant information that may be miscontained in controllable latent state z^+. We conclude that both of these two designs are absolutely necessary for imitation learning in environments with complex observations.

5.3. Could SeMAIL perform well with distractors that never appear in expert observations?

To test SeMAIL's ability by solving tasks with distractors that the expert has never seen before, we train the agent and expert in environments with non-overlap video backgrounds for Walker Run, Cheetah Run, and Finger Spin tasks. Due to the different backgrounds, the discriminator tends to distinguish the agent's observations from the expert's, which leads to low rewards and failure to learn expert behaviors. We record the mean and standard error of the maximum expected episodic returns over all the methods in tests and scale the values such that 0 represents the random agent performance and 1 represents the expert performance. The results in Table 1 show V-MAIL and DA-DAC fail to solve the task with very low performances. DisentanGAIL performs well in the Finger Spin task but poorly in Walker and Cheetah tasks. It indicates that regularizing the mutual information between the agent and expert observations can mitigate the misdistinguishing problem to some extent in this experiment. The reason that SeMAIL (No BoR) unexpectedly got an outstanding performance in the Walker task may be its task model remains more observation In-
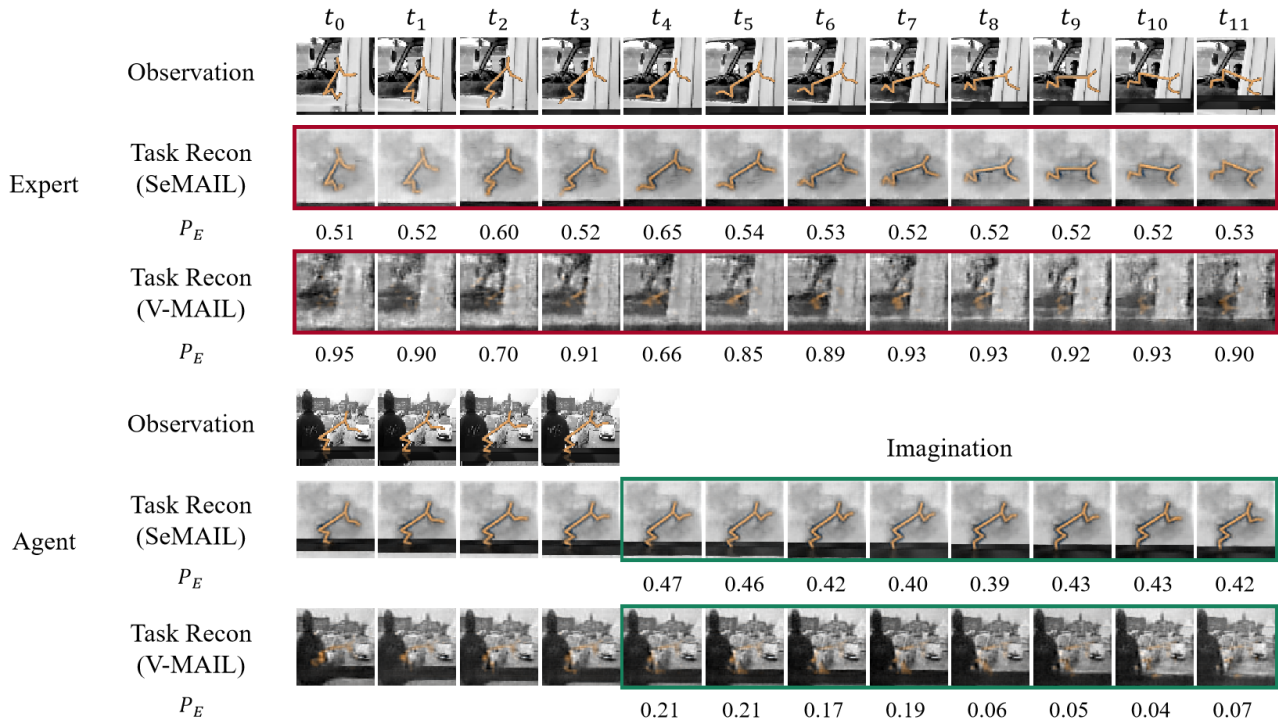
7

Figure 5. The adversarial imitation learning process of SeMAIL and V-MAIL. In latent space, both of SeMAIL and V-MAIL imagine the trajectory based on the observations of $t_0 \sim t_3$ and imitate the expert behavior from all expert demonstrations. The reconstructions of the task model for the expert observation and the agent imagination are presented in red and green boxes, respectively. The value P_E denotes the probability that discriminator D_ψ predicts agent behaviors as expert behaviors, which is closer to 0.5 will be better.

Table 1. Performance on tasks with non-overlap video backgrounds from expert observations, scaled by the expert and random agent returns.

METHOD	WALKER RUN	CHEETAH RUN	FINGER SPIN
V-MAIL	0.16 ± 0.01	0.17 ± 0.02	0.02 ± 0.01
DA-DAC	0.13 ± 0.02	0.19 ± 0.02	0.09 ± 0.03
DISENTANGAIL	0.27 ± 0.04	0.19 ± 0.03	$\mathbf{1.47 \pm 0.07}$
SEMAIL (NO ACT)	0.10 ± 0.01	0.31 ± 0.04	0.07 ± 0.02
SEMAIL (NO BOR)	$\mathbf{0.84 \pm 0.01}$	0.38 ± 0.03	0.75 ± 0.08
SEMAIL	0.38 ± 0.04	$\mathbf{0.86 \pm 0.04}$	0.82 ± 0.05

formation, which could benefit some different background situations. Although all the methods can not reach expert performance, our method still has a substantial advantage over other baselines.

5.4. How well does the task-relevant space imagination benefit adversarial imitation learning?

To answer this question, we visualize the reconstruction of Cheetah Run's task information for expert observations and

agent imaginations by SeMAIL and V-MAIL in Figure 5. The agent and the expert are trained in environments with the same task but with non-overlap video backgrounds. In expert observations, SeMAIL only extracts the cheetah's body as task information, while V-MAIL reconstructs all the information from the original image. In agent observations, SeMAIL imagines its behavior in the task-relevant latent space $\mathcal{Z}^+$, while V-MAIL imagines in the latent space containing all the information. Based on the reconstruction of the imagined data, we can see SeMAIL filter out irrelevant information while V-MAIL largely retains it. To evaluate the learning ability in the adversarial imitation learning process, we use the probability of whether the discriminator can distinguish behaviors from the agent or the expert. For agent imagination using V-MAIL, the discriminator gives a near-zero probability value since it carries all observed information which leads to distinguish from the expert (as the example mentioned in Figure 1). The agent of V-MAIL nearly can not receive positive training signals from the discriminator and underperforms in the task as a result. The discriminator of SeMAIL scores a probability close to 0.5 based on task-relevant information, which guides the agent's policy learning toward the expert.

6. Conclusion

In model-based imitation learning, irrelevant information will cause the discriminator to have deceptive biases on the reward. We introduce the Action-conditioned Transition assumption to model task-relevant and distractor dynamics separately, and we name this new approach Separated Model-based Adversarial Imitation Learning (SeMAIL). SeMAIL extracts task-correlated features from the environment, which helps mitigate the problem of current MBIL methods struggling to learn in tasks with complex distractors. We provide a theoretical demonstration that the performance gap between the expert and the agent can be upper-bounded in task-relevant space. We verify the performance of SeMAIL on six frequently-used continuous control tasks, with complex distractors in observations. Further, we design ablation studies and visualize the adversarial imitation learning process to display the contribution of the key components. We conclude that SeMAIL can provide a relatively clean task-relevant latent space for the adversarial imitation learning process and largely improve performance on VIL tasks with complex observations.

In our experiments, we claim that irrelevant information in most environments is a distraction to the agent's action decision and try to eliminate it. In particular cases, the irrelevant distractor may become task-relevant in a certain condition and influence the agent's decision-making. In this case, we can add a short period of the particular distractor states in the policy function and train the agent. We will consider this improvement in future work. Another further improvement is training the task-irrelevant models of the expert and the agent separately, which may benefit the scenario where distractors in their observations are majorly different.

Acknowledgements

References

Baram, N., Anschel, O., and Mannor, S. Model-based adversarial imitation learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016.

Barber, D. and Agakov, F. V. The IM algorithm: a variational approach to information maximization. In *Advances in Neural Information Processing Systems 16 :Neural Information Processing Systems 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada*, pp. 201–208. MIT Press, 2003.

Bharadhwaj, H., Babaeizadeh, M., Erhan, D., and Levine, S. Information prioritization through empowerment in visual model-based RL. In *10th International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI gym. *CoRR*, abs/1606.01540, 2016.

Cetin, E. and Çeliktutan, O. Domain-robust visual imitation learning with mutual information constraints. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Cohen, S., Amos, B., Deisenroth, M. P., Henaff, M., Vinitsky, E., and Yarats, D. Imitation learning from pixel observations for continuous control. 2021.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013.

Deisenroth, M. P. and Rasmussen, C. E. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 465–472. Omnipress, 2011.

Du, S. S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudík, M., and Langford, J. Provably efficient RL with rich observations via latent state decoding. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1665–1674. PMLR, 2019.

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27):6435–6467, 2021.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 49–58. JMLR.org, 2016.

Finn, C., Yu, T., Zhang, T., Abbeel, P., and Levine, S. One-shot visual imitation learning via meta-learning. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pp. 357–368. PMLR, 2017.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adverserial inverse reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Fu, X., Yang, G., Agrawal, P., and Jaakkola, T. S. Learning task informed abstractions. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3480–3491. PMLR, 2021.

Ghasemipour, S. K. S., Zemel, R. S., and Gu, S. A divergence minimization perspective on imitation learning methods. In *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1259–1277. PMLR, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5767–5777, 2017.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., and Dragan, A. D. Inverse reward design. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6765–6774, 2017.

Hafner, D., Lillicrap, T. P., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019.

Hafner, D., Lillicrap, T. P., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Hafner, D., Lillicrap, T. P., Norouzi, M., and Ba, J. Mastering Atari with discrete world models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Haldar, S., Mathur, V., Yarats, D., and Pinto, L. Watch and match: Supercharging imitation with regularized optimal transport. *CoRR*, abs/2206.15469, 2022. doi: 10.48550/arXiv.2206.15469.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2980–2988. IEEE Computer Society, 2017.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4565–4573, 2016.

Hu, A., Corrado, G., Griffiths, N., Murez, Z., Gurau, C., Yeo, H., Kendall, A., Cipolla, R., and Shotton, J. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

Kostrikov, I., Agrawal, K. K., Dwibedi, D., Levine, S., and Tompson, J. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Liu, M., He, T., Zhang, W., Shuicheng, Y., and Xu, Z. Visual imitation learning with patch rewards. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Found. Trends Robotics*, 7(1-2):1–179, 2018. doi: 10.1561/2300000053.

Pan, M., Zhu, X., Wang, Y., and Yang, X. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23178–23191. Curran Associates, Inc., 2022.

Pathak, D., Mahmoudieh, P., Luo, G., Agrawal, P., Chen, D., Shentu, Y., Shelhamer, E., Malik, J., Efros, A. A., and Darrell, T. Zero-shot visual imitation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Rafailov, R., Yu, T., Rajeswaran, A., and Finn, C. Visual adversarial imitation learning using variational models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 3016–3028, 2021.

Ross, S., Gordon, G. J., and Bagnell, D. 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, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pp. 627–635. JMLR.org, 2011.

Seo, Y., Lee, K., James, S. L., and Abbeel, P. Reinforcement learning with action-free pre-training from videos. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19561–19579. PMLR, 2022.

Stadie, B. C., Abbeel, P., and Sutskever, I. Third person imitation learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Sutton, R. S. and Barto, A. G. Time-derivative models of Pavlovian reinforcement. 1990.

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T. P., and Riedmiller, M. A. DeepMind control suite. *CoRR*, abs/1801.00690, 2018.

Tishby, N. and Slonim, N. Data clustering by Markovian relaxation and the information bottleneck method. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems 2000, Denver, CO, USA*, pp. 640–646. MIT Press, 2000.

Torabi, F., Warnell, G., and Stone, P. Imitation learning from video by leveraging proprioception. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 3585–3591. ijcai.org, 2019. doi: 10.24963/ijcai.2019/497.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Wang, T., Du, S., Torralba, A., Isola, P., Zhang, A., and Tian, Y. Denoised MDPs: Learning world models better than the world itself. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22591–22612. PMLR, 2022.

Xiao, T., Radosavovic, I., Darrell, T., and Malik, J. Masked visual pre-training for motor control. *CoRR*, abs/2203.06173, 2022. doi: 10.48550/arXiv.2203.06173.

Yarats, D., Kostrikov, I., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Young, S., Gandhi, D., Tulsiani, S., Gupta, A., Abbeel, P., and Pinto, L. Visual imitation made easy. In *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, volume 155 of *Proceedings of Machine Learning Research*, pp. 1992–2005. PMLR, 2020.

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Zolna, K., Reed, S. E., Novikov, A., Colmenarejo, S. G., Budden, D., Cabi, S., Denil, M., de Freitas, N., and Wang, Z. Task-relevant adversarial imitation learning. In *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, volume 155 of *Proceedings of Machine Learning Research*, pp. 247–263. PMLR, 2020.

A. Proof

A.1. Proof of Theorem 4.1

Theorem A.1. *Consider a POMDP $\mathcal{M}$ with high-dimensional inputs such as images. Let s_t be the ground-truth state and z_t be the latent representation of the whole observation o_t. z_t^+ and z_t^- represent for the task-relevant and irrelevant component, respectively, which meets the condition of $p(s_t|z_t, a_t) = p(s_t|z_t^+, a_t)p(s_t|z_t^-)$ based on the AcT assumption mentioned in Section 4.1. Then, we have*

$$\mathbb{D}_f(\rho_\mathcal{M}^\pi(o,a)||\rho_\mathcal{M}^E(o,a)) \leq \mathbb{D}_f(\rho_\mathcal{M}^\pi(s,a)||\rho_\mathcal{M}^E(s,a)) \leq \mathbb{D}_f(\rho_\mathcal{M}^\pi(z^+,a)||\rho_\mathcal{M}^E(z^+,a)) \tag{8}$$

where $\mathbb{D}_f$ is a generic f-divergence.

Proof. We get the result that $\mathbb{D}_f(\rho_\mathcal{M}^\pi(o,a)||\rho_\mathcal{M}^E(o,a)) \leq \mathbb{D}_f(\rho_\mathcal{M}^\pi(s,a)||\rho_\mathcal{M}^E(s,a)) \leq \mathbb{D}_f(\rho_\mathcal{M}^\pi(z,a)||\rho_\mathcal{M}^E(z,a))$ from Theorem 1 which is proved in Rafael Rafailov's work (2021). Next, we need to prove $\mathbb{D}_f(\rho_\mathcal{M}^\pi(z,a)||\rho_\mathcal{M}^E(z,a)) \leq \mathbb{D}_f(\rho_\mathcal{M}^\pi(z^+,a)||\rho_\mathcal{M}^E(z^+,a))$, which means that the gap between the agent and the expert occupancy measures upper-bounded in the latent state space $\mathcal{Z}$ can also be upper-bounded in the task-relevant state space $\mathcal{Z}^+$.

$$\mathbb{D}_f(\rho_\mathcal{M}^\pi(z,a)||\rho_\mathcal{M}^E(z,a)) = \mathbb{E}_{z,a\sim\rho_\mathcal{M}^E(z,a)}\left[f\left(\frac{\rho_\mathcal{M}^\pi(z,a)}{\rho_\mathcal{M}^E(z,a)}\right)\right] \tag{9}$$

$$= \mathbb{E}_{z^+,z^-,a\sim\rho_\mathcal{M}^E(z,a)}\left[f\left(\frac{\rho_\mathcal{M}^\pi(z^+,a)P(z^-|z^+)}{\rho_\mathcal{M}^E(z^+,a)P(z^-|z^+)}\right)\right] \tag{10}$$

$$= \mathbb{E}_{z^+,z^-,a\sim\rho_\mathcal{M}^E(z,a)}\left[f\left(\frac{\rho_\mathcal{M}^\pi(z^+,a)}{\rho_\mathcal{M}^E(z^+,a)}\right)\right] \tag{11}$$

$$\leq \mathbb{E}_{z^+,a\sim\rho_\mathcal{M}^E(z^+,a)}\left[f\left(\frac{\rho_\mathcal{M}^\pi(z^+,a)}{\rho_\mathcal{M}^E(z^+,a)}\right)\right] \tag{12}$$

$$= \mathbb{D}_f(\rho_\mathcal{M}^\pi(z^+,a)||\rho_\mathcal{M}^E(z^+,a)) \tag{13}$$

The equality (10) follows the fact that $\rho_\mathcal{M}(z,a) = \rho_M(z^+,z^-,a) = \rho_\mathcal{M}(z^+,a)P(z^-|z^+,a) = \rho_\mathcal{M}(z^+,a)P(z^-|z^+)$, which means the task-relevant and -irrelevant parts can be partitioned, as mentioned in Section 4.1. The inequality (12) is deflated with the help of our AcT assumption, which assumes that the forward dynamics in $\mathcal{Z}$ can be decoupled into two independent forward dynamics in $\mathcal{Z}^+$ and $\mathcal{Z}^-$.

□

B. Derivations

Previous model-based RL studies (Hafner et al., 2020; 2021) use the information bottleneck objective (Tishby & Slonim, 2000) to encourage model states to predict observations and rewards while limiting the capacity of information that contained in states. We remove the reward prediction in the objective for model-based imitation learning as follows:

$$\max \mathrm{I}(o_{1:T}, z_{1:T}|a_{1:T}) - \beta\mathrm{I}(i_{1:T}, z_{1:T}|a_{1:T}) \tag{14}$$

Here, i_t is indices of the dataset such that $p(o_t|i_t) = \delta(o_t - o_t')$. The first term can be simplified using the mutual information definition and the non-negativity of the KL-divergence.

$$I(o_{1:T}, z_{1:T}|a_{1:T}) = \mathbb{E}_{p(o_{1:T},z_{1:T},a_{1:T})}\left[\ln p(o_{1:T}|z_{1:T}, a_{1:T}) - \ln p(o_{1:T}|a_{1:T})\right] \tag{15}$$

$$\overset{+}{=} \mathbb{E}_{p(o_{1:T},z_{1:T},a_{1:T})}\left[\ln p(o_{1:T}|z_{1:T}, a_{1:T})\right] \tag{16}$$

$$\geq \mathbb{E}_{p(o_{1:T},z_{1:T},a_{1:T})}\left[\ln p(o_{1:T}|z_{1:T}, a_{1:T})\right] - \mathbb{D}_{\text{KL}}\Big(p(o_{1:T}|z_{1:T}, a_{1:T})||\prod_{t=1}^{T} q(o_t|z_t)\Big) \tag{17}$$

$$= \mathbb{E}_{q(z_{1:T}|o_{1:T},a_{1:T})}\left[\sum_{t=1}^{T}\ln q(o_t|z_t)\right] \tag{18}$$

$$= \sum_{t=1}^{T}\left[\mathbb{E}_{q(z_t^+|o_{1:t},a_{1:t-1})q(z_t^-|o_{1:t})}\ln q(o_t|z_t^+, z_t^-)\right] \tag{19}$$

The second term in the equality (15) can be regarded as constant for the observed data. The equality (19) is obtained by the fact that the representation of observation z_t can be decoupled into the task-relevant part z_t^+ and irrelevant part z_t^-.

For the second term, we obtain the upper bound of it with the non-negativity of the KL-divergence and the AcT assumption below:

$$I(z_{1:T}, i_{1:T}|a_{1:T}) = \mathbb{E}_{p(o_{1:T},z_{1:T},a_{1:T},i_{1:T})}\left[\sum_{t=1}^{T}\ln p(z_t|z_{t-1}, a_{t-1}, i_t) - \ln p(z_t|z_{t-1}, a_{t-1})\right] \tag{20}$$

$$\leq \mathbb{E}_{q(z_{1:T}|o_{1:T},a_{1:T-1})}\left[\sum_{t=1}^{T}\ln \frac{q(z_t|o_t, z_{t-1}, a_{t-1})}{p(z_t|a_{t-1}, z_{t-1})}\right] \tag{21}$$

$$= \mathbb{E}_{q(z_{1:t-1}|o_{1:t-1},a_{1:t-2})}\left[\sum_{t=1}^{T}\mathbb{E}_{q(z_t|o_t,z_{t-1},a_{t-1})}\ln \frac{q(z_t|o_t, z_{t-1}, a_{t-1})}{p(z_t|a_{t-1}, z_{t-1})}\right] \tag{22}$$

$$= \mathbb{E}_{q(z_{1:t-1}|o_{1:t-1},a_{1:t-2})}\left[\sum_{t=1}^{T}\mathbb{E}_{q(z_t|o_t,z_{t-1},a_{t-1})}\ln \frac{q(z_t^+|o_t, z_{t-1}^+, a_{t-1})q(z_t^-|o_t, z_{t-1}^-)}{p(z_t^+|z_{t-1}^+, a_{t-1})p(z_t^-|z_{t-1}^-)}\right] \tag{23}$$

$$= \mathbb{E}_{q(z_{1:t-1}^+|o_{1:t-1},a_{1:t-2})}\left[\sum_{t=1}^{T}\mathbb{E}_{q(z_t^+|o_t,z_{t-1}^+,a_{t-1})}\ln \frac{q(z_t^+|o_t, z_{t-1}^+, a_{t-1})}{p(z_t^+|z_{t-1}^+, a_{t-1})}\right]$$

$$+ \mathbb{E}_{q(z_{1:t-1}^-|o_{1:t-1})}\left[\sum_{t=1}^{T}\mathbb{E}_{q(z_t^-|o_t,z_{t-1}^-)}\ln \frac{q(z_t^-|o_t, z_{t-1}^-)}{p(z_t^-|z_{t-1}^-)}\right] \tag{24}$$

$$= \sum_{t=1}^{T}\Big(\mathbb{E}_{q(z_{t-1}^+|o_{1:t-1},a_{1:t-2})}\left[\mathbb{D}_{\text{KL}}(q(z_t^+|o_t, z_{t-1}^+, a_{t-1})||p(z_t^+|z_{t-1}^+, a_{t-1}))\right]$$

$$+ \mathbb{E}_{q(z_{t-1}^-|o_{1:t-1})}\left[\mathbb{D}_{\text{KL}}(q(z_t^-|o_t, z_{t-1}^-)||p(z_t^-|z_{t-1}^-))\right]\Big) \tag{25}$$

In practice, we use two pairs of observation decoders q_{ϕ^+} and q_{ϕ^-}, forward dynamics models p_{θ^+} and p_{θ^-}, and variational posterior models q_{ψ^+} and q_{ψ^-} for the task-relevant and irrelevant branches respectively. We obtain the final objective of the separated models to be optimized as follows:

$$\max_{\substack{\theta^+,\theta^-,\psi^+,\psi^- \\ \phi^+,\phi^-,\tilde{\phi}}} \mathbb{E}_{(o_\tau,a_\tau)\sim\mathcal{B}_\pi\cup\mathcal{B}_E}\Bigg[\sum_{t=1}^{T}\Big(\mathbb{E}_{q(z_{t-1}^-|o_{1:t-1})}[-\mathbb{D}_{\text{KL}}\left(q_{\psi^-}(z_t^-|o_t, z_{t-1}^-)||p_{\theta^-}(z_t^-|z_{t-1}^-)\right)]$$

$$+\mathbb{E}_{q(z_{t-1}^+|o_{1:t-1},a_{1:t-2})}[-\mathbb{D}_{\text{KL}}\left(q_{\psi^+}(z_t^+|o_t, z_{t-1}^+, a_{t-1})||p_{\theta^+}(z_t^+|z_{t-1}^+, a_{t-1})\right)]$$

$$+\mathbb{E}_{q(z_t^+|o_{1:t},a_{1:t-1})q(z_t^-|o_{1:t})}[\ln q_{\phi^+,\phi^-}(o_t|z_t^+, z_t^-) + \lambda\ln q_{\tilde{\phi}}(o_t|z_t^-)]\Big)\Bigg] \tag{26}$$

where we add the background-only reconstruction loss mentioned in Section 4.2.

Figure 6. The environments used in our experiments: Walker Walk, Walker Run, Cheetah Run, Finger Spin, Hopper Hop, and Car Racing, where the first five replace backgrounds with natural videos except for Car Racing.

C. Implementation Details

C.1. Environments and Tasks

DeepMind Control Suite (DMC) (Tassa et al., 2018) is a set of reinforcement learning environments that includes a range of tasks, such as locomotion, manipulation, and navigation. We choose five locomotion tasks from DMC and replace the background wall with the grayscaled videos in the "driving car" class from the kinematic dataset (Kay et al., 2017). These natural video backgrounds contain complex task-irrelevant information. The Car Racing environment from OpenAI Gym (Brockman et al., 2016) is a classic control task from pixels. The agent must learn to control the speed, steering, and other parameters to maximize its cumulative rewards. We cut out the bottom part of the frames to avoid the agent learning directly from the reward signals. The size of image observation in all environments is $64 \times 64 \times 3$. The example observations of these environments are shown in Figure 6.

In the first part of our experiments, we test our method SeMAIL and the baselines on all six environments. On the five environments of DMC, the agent and the expert observations background walls are replaced with the frames from the same videos. We design this task to verify SeMAIL's learning ability in environments with complex and noisy observations. We adopt the original image input for the Car Racing environment because it already contains much irrelevant information as time-varying blocks in observations.

In the following experiment, we use Walker Run, Cheetah Run, and Finger Spin tasks to test SeMAIL's ability with backgrounds not appeared in expert observations. To achieve this, we design two non-overlap environments for each task. We train the expert in the first eight videos under the "driving car" class and collect the demonstration, then we train the agent in the last eight videos in the same class with the expert demonstration.

C.2. Demonstration Data

To make sure agent actions can be inferred from image inputs with distractors, we train an expert from scratch for each task and obtain the demonstration data. We train the expert with TIA (Fu et al., 2021) for locomotion tasks from DMC and train it with Dreamer (Hafner et al., 2020) for the Car Racing task. We prepare for all the methods with 10 expert demonstrations in each task.

C.3. Pseudo Code

The pseudo-code of our proposed SeMAIL is provided in Algorithm 1.

C.4. Networks and Hyperparameters

Implementation. We implement the proposed algorithm with TensorFlow 2 and run all the experiments on NVIDIA RTX 3090 for about 1000 GPU hours. We use the recurrent state space model (Hafner et al., 2019) for the forward dynamics and the posterior encoder. The hidden sizes for the deterministic part and stochastic part are 200 and 30. We adopt the convolutional encoder and decoder used in TIA (Fu et al., 2021). The size of all dense layers is 300, and the activation function is ELU. We use ADAM optimizer to train the network with batches of 64 sequences of length 50. The learning rate for the task and background model is 6e-5, and for the action net, value net, and discriminator is 8e-5. We clip gradient norms to 100 to stabilize the training process. To prevent training a too-strong discriminator, we add a gradient penalty term (Gulrajani et al., 2017) on the discriminator loss and set the value of weight as 1.0. The codes of SeMAIL are released in

---

Algorithm 1 Training Procedure of SeMAIL

    Input: Policy replay buffer $\mathcal{B}_\pi$, Expert demonstrations $\mathcal{B}_E$

    Initialize forward dynamics model $p_{\theta^+}, p_{\theta^-}$, posterior encoder q_{ψ^+}, q_{ψ^-}, observation decoder $q_{\phi^+}, q_{\phi^-}, q_{\tilde{\phi}}$, policy π.

    for each time step $t = 1 \cdots T$ **do**

        // Rollout trajectories

        Infer the task-relevant latent state $z_t^+ \sim q_{\psi^+}(\cdot|o_t, z_{t-1}^+, a_{t-1})$

        Sample action from policy $a_t \sim \pi(\cdot|z_t^+)$

        Execute action and get the next observation $o_{t+1} \leftarrow \text{env.step}(a_t)$

    end for

    Add samples into the replay buffer $\mathcal{B}_\pi \leftarrow \mathcal{B}_\pi \cup \{(o_t, a_t)_{t=1}^T\}$

    for training iteration $i = 1 \cdots \text{It}$ **do**

        // Learn separated models

        Sample minibatch $(o_{1:T}, a_{1:T-1})_{1:b}$ from the union buffer $\mathcal{B}_\pi \cup \mathcal{B}_E$

        Update the forward dynamics model $p_{\theta^+}, p_{\theta^-}$ and the posterior encoder q_{ψ^+}, q_{ψ^-} with Equation (3)

        Update the observation decoder $q_{\phi^+}, q_{\phi^-}, q_{\tilde{\phi}}$ with Equation (4)

        // Optimize policy

        Imagine the task-relevant latent states $z_{1:H}^{\pi^+}$ by policy π using the forward dynamics model p_{θ^+}

        Sample expert's trajectories $(o_{1:T}^E, a_{1:T-1}^E)$ from demonstration buffer $\mathcal{B}_E$

        Infer the task-relevant latent states $z_{1:T}^{E+}$ using the posterior encoder q_{ψ^+}

        Train the discriminator on the data $(z_{1:H-1}^{\pi^+}, a_{1:H-1}^\pi), (z_{1:T-1}^{E+}, a_{1:T-1}^E)$ using Equation (7)

        Update the policy π to imitate the expert's behavior using Equation (8)

    end for

---

https://github.com/yixiaoshenghua/SeMAIL.

Environment Hyperparameters. For Walker Walk, Walker Run, Cheetah Run, Finger Spin, Hopper Hop, and Car Racing tasks, the values of background-only reconstruction λ are 1.5, 0.25, 2, 1.5, 2, and 1, respectively. To encourage exploration, we add $\mathcal{N}(0, 0.3)$ noise on the output actions for locomotion tasks and $\mathcal{N}(0, 0.1)$ noise for Car Racing as used in (Rafailov et al., 2021). The imagination horizon H for locomotion tasks is 15, and for Car Racing is 10.

Training Details. To make this a fair comparison, all environments and network parameters are kept the same over SeMAIL and the compared algorithms. We initialize the dataset with 5 randomly collected episodes and train 100 iterations after collecting one episode in environments. We keep the action repeat times as 2 and set the discounting factor as 0.99 for all tasks. On five locomotion tasks, we adopt the official implementation for DisentanGAIL (Cetin & Çeliktutan, 2021), which obtains pseudo rewards from the discriminator on the agent observations and learns policy based on the raw state observations.

D. Extra Experimental Results

D.1. Quantitative Results

In Table 2, we show the quantitative results of the experiments in Section 5.1. SeMAIL significantly outperforms the baselines in almost all tasks and achieves near-expert performance. Without the AcT assumption or removing the BoR loss, the performances of SeMAIL in locomotion tasks significantly decrease in either case. On the Car Racing task, the performance of SeMAIL (No BoR) outperforms SeMAIL, and V-MAIL achieves better performance than on other tasks. It is because the background of the car racing environment is relatively simple, and the task-relevant part takes up the large proportion. DisentanGAIL performs well on Finger Spin but fails on other tasks due to its agent actions change little in observations, which regularizing the mutual information between the expert and the agent data can help solve this task.

D.2. Additional Visualization

We use the t-SNE plot (van der Maaten & Hinton, 2008) to visualize training result through the agent's and expert's observation embeddings of SeMAIL and V-MAIL in Figure 7. We randomly select images from observations of the initial stage, the learned policy, and the expert, then generate the latent representation $z^+ \sim q_\phi^+(z^+|o)$ from SeMAIL and

Table 2. Performance on six visual control tasks. We present the mean and std of final performance by running 10 trajectories over 4 seeds for SeMAIL and the baselines.

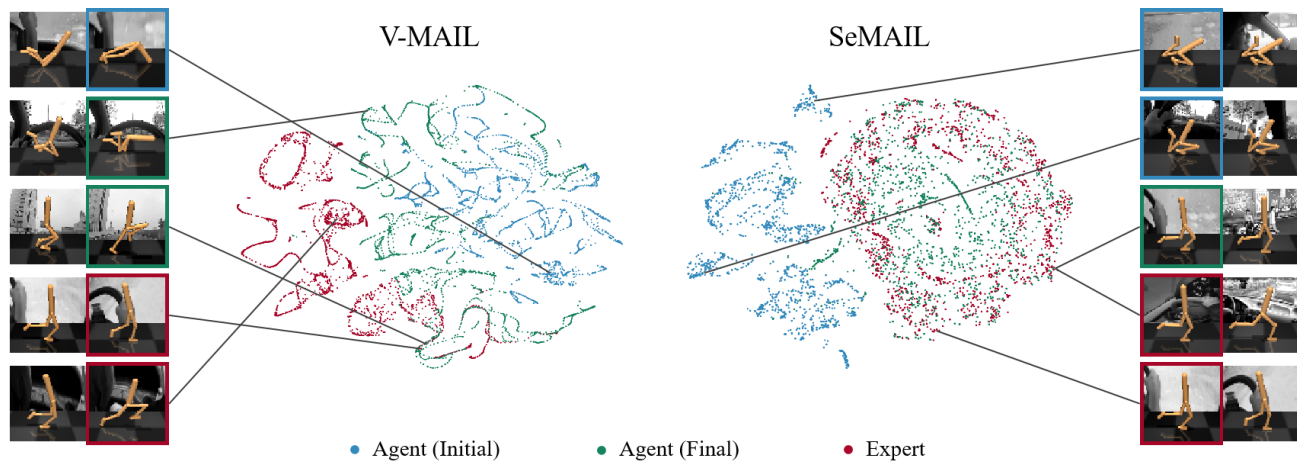| METHOD	WALKER WALK	WALKER RUN	CHEETAH RUN	FINGER SPIN	HOPPER HOP	CAR RACING
EXPERT	964.5 ± 14.7	539.5 ± 6.5	629.5 ± 44.6	332.0 ± 24.4	263.2 ± 7.4	936.3 ± 5.7
V-MAIL	314.9 ± 295.6	155.1 ± 73.9	74.6 ± 48.9	3.3 ± 4.8	0.2 ± 1.0	537.8 ± 228.4
DA-DAC	25.4 ± 13.3	27.7 ± 21.2	32.9 ± 17.0	0.1 ± 0.6	0.0 ± 0.1	-57.9 ± 38.7
DISENTANGAIL	61.0 ± 6.6	68.7 ± 13.4	36.6 ± 17.3	133.2 ± 77.9	1.0 ± 1.1	-62.0 ± 20.0
SEMAIL (NO ACT)	124.9 ± 14.1	100.0 ± 7.4	71.9 ± 4.4	30.5 ± 4.3	66.3 ± 8.5	24.1 ± 36.2
SEMAIL (NO BOR)	93.7 ± 15.5	337.5 ± 14.7	72.9 ± 5.8	23.4 ± 5.4	19.1 ± 2.1	$\mathbf{942.9 \pm 1.6}$
SEMAIL	$\mathbf{900.1 \pm 57.6}$	$\mathbf{463.4 \pm 42.7}$	$\mathbf{217.9 \pm 125.8}$	$\mathbf{161.1 \pm 68.7}$	$\mathbf{90.5 \pm 75.6}$	901.2 ± 19.8

Figure 7. The t-SNE plot of data representations of the initial agent, the final learned agent, and the expert. Compared with the V-MAIL plot on the left, the embedding of our method SeMAIL shows that the learned agent behavior blends well with the expert, which indicates well-trained by eliminating distractors.

$z \sim q(z|o)$ from V-MAIL. Since misguided by complex distractors, the embeddings cluster of V-MAIL shows a large gap between the trained agent and the expert. In SeMAIL, the sample embeddings almost cover the representation space of the expert observations, which indicates a well-trained result through our distractor-eliminating policy. It confirms that SeMAIL is qualified to solve real-world cases which may contain complex distractors.

E. Additional Discussions

E.1. Pre-trained Visual Encoders

Equipping visual RL with a pre-trained encoder is an exciting direction like recent research (Xiao et al., 2022; Seo et al., 2022). In this paper, we are concerned about the practical problem in visual imitation learning (VIL) - how to extract task-relevant information from noisy observations without access to ground-truth rewards. To solve this, a good alignment between the agent and the expert behaviors (by separating the dynamics of task-relevant and -irrelevant parts) is a more important contributor than a good representation of observations (by pre-training or data augmentation). The compared method in our experiments, DA-DAC, only considers obtaining a good representation via augmenting the observed data, and its agent finally receives the wrong guidance from the expert and fails to complete the tasks (see the quantitative results in Table 2). We leave the pre-training approach combined with the SeMAIL framework for future exploration.

Table 3. Performance on five DMC tasks. We present the mean and std of 10 trajectories' return for SeMAIL and SeMAIL (No Exp-M).

METHOD	WALKER WALK	WALKER RUN	CHEETAH RUN	FINGER SPIN	HOPPER HOP
SEMAIL	900.1 ± 57.6	463.4 ± 42.7	217.9 ± 125.8	161.1 ± 68.7	90.5 ± 75.6
SEMAIL (NO EXP-M)	62.8 ± 11.5	63.7 ± 17.4	15.1 ± 13.3	3.0 ± 4.5	1.2 ± 2.5

E.2. The Rationality of POMDP Assumption

POMDP (Kaelbling et al., 1998) addresses that the agent can not directly access the ground truth states but only infer the belief states from observations. In previous RL work (Rafailov et al., 2021; Bharadhwaj et al., 2022) using the similar environments with us, they adopted POMDP as their basic environmental assumption. We follow their assumption because we aim to solve imitation learning tasks with noisy observations, which have some common ground with theirs. As detailed in Appendix C.4, our transition model is built upon a recurrent state space model, which will generate belief hidden states from historical rollouts for policy action outputs. BMDP (Du et al., 2019) highlights the block structure that each context determines its generating state uniquely, which can be regarded as an alternative formulation of our problem. We highly encourage researchers with an interest in BMDP to consult the original paper.

E.3. The Practice of the AcT Assumption

The AcT assumption claims that the forward dynamics $p(z_t|z_{t-1}, a_{t-1})$ can be **independently** decomposed into the product of task-relevant dynamics $p(z_t^+ \mid z_{t-1}^+, a_{t-1})$ and -irrelevant dynamics $p(z_t^- | z_{t-1}^-)$, and each one is implemented by a stochastic probabilistic model. Recent studies, including Denoised MDPs (Wang et al., 2022) and Iso-Dream (Pan et al., 2022), consider modeling the environmental dynamics using controllable information and pre-defined rewards. Denoised MDPs categorize information with controllablity and reward relativity. Iso-Dream posits that the latent states can be separated into controllable and noncontrollable part, while the agent makes decisions influenced by future noncontrollable states and subsequently affecting controllable states. Motivated by experimental observations and in light of the computational savings afforded by the independence hypothesis, we propose the AcT assumption. In contrast to previous work, the AcT assumption is stronger, positing that task-relevant and -irrelevant dynamics are independent. During the initial phase of training, due to the limited data collected by the agent, the separated models primarily rely on expert data for training. Because expert policies are optimal and focus only on task-relevant information to make actions unaffected by distractors, task-relevant and -irrelevant models can be considered independent at this stage. As training progress progresses and the agent's policy approaches optimality, the agent becomes more resistant to irrelevant information interference, which supports the independence assumption in AcT. To testify the necessity of expert data for the AcT assumption in model learning of SeMAIL, we conduct an additional experiment in which separated models are trained without expert data while all other settings remained consistent. This method is denoted as SeMAIL (No Exp-M) in Table 3. SeMAIL (No Exp-M) exhibits a significant performance decline, indicating the crucial role of expert data in model training for imitation learning under the AcT assumption.

In the implementation of SeMAIL based on the AcT assumption, we utilize a unimodal Gaussian distribution as the policy output. The environments and expert's data in our experiments all feature an optimal unimodal action distribution. This implicit environmental assumption which fits SeMAIL's policy output, may contribute to SeMAIL's near-expert performance shown in Table 2. In complex environments with multimodal optimal behavior distributions, direct using current SeMAIL's policy may result in performance degradation. A potential solution is to modify the policy's action output to a mixture of Gaussian distributions and introduce an additional loss, which aligns the policy's action output with the expert's action distribution without explicitly altering the AcT assumptions or other components of SeMAIL.

In particular cases, a previous distractor could turn into a task-relevant factor in a certain condition and influence the agent's further decision-making. In this case, we can add a short period of the particular distractor states in the policy function and train the agent. We will consider these extensions in our future work.

E.4. Modifications of Non-IL Methods in Noisy Observations

In Section 2, we have clarified that the Non-IL methods in noisy observations heavily rely on using environmental rewards to exclude distractions. Our concern is how to learn a good policy under noisy observations (e.g., natural background videos) in imitation learning, which is free from reward-design issues. To our knowledge, our work is the first to consider this setting

Table 4. Performance on three DMC tasks. We present the mean and std of 10 trajectories' return for random policy and Non-IL method with pseudo reward.

METHOD	WALKER RUN	FINGER SPIN	HOPPER HOP
RANDOM POLICY	29.0 ± 11.8	2.1 ± 2.8	0.0 ± 0.0
WITH PSEUDO REWARD	50.4 ± 17.2	3.3 ± 5.0	0.7 ± 2.2

in IL. When designing the experiments, we attempted to modify related methods to adapt the setting of our work by utilizing the estimated pseudo rewards given by the discriminator to separate task-relevant and -irrelevant information. However, they failed to solve the tasks (the "with pseudo reward" row in Table 4). Using ill-estimated rewards can result in biased task-related representations and further mistaken estimation of rewards, which leads to a vicious cycle.