Guaranteed Discovery of Control-Endogenous Latent States with Multi-Step Inverse Models

Anonymous authors
Paper under double-blind review

Abstract

In many sequential decision-making tasks, the agent is not able to model the full complexity of the world, which consists of multitudes of relevant and irrelevant information. For example, a person walking along a city street who tries to model all aspects of the world would quickly be overwhelmed by a multitude of shops, cars, and people moving in and out of view, each following their own complex and inscrutable dynamics. Is it possible to turn the agent's firehose of sensory information into a minimal latent state that is both necessary and sufficient for an agent to successfully act in the world? We formulate this question concretely, and propose the Agent Control-Endogenous State Discovery algorithm (`AC-State`), which has theoretical guarantees and is practically demonstrated to discover the *minimal control-endogenous latent state* which contains all of the information necessary for controlling the agent, while fully discarding all irrelevant information. This algorithm consists of a multi-step inverse model (predicting actions from distant observations) with an information bottleneck. `AC-State` enables localization, exploration, and navigation without reward or demonstrations. We demonstrate the discovery of the control-endogenous latent state in three domains: localizing a robot arm with distractions (e.g., changing lighting conditions and background), exploring a maze alongside other agents, and navigating in the Matterport house simulator.

1 Introduction

In many real-world systems, the observation space is generated by multiple complex and sparsely interacting subsystems. The state that is necessary for controlling an agent depends only on a small fraction of the information in the observation space. For example, consider a world consisting of a robot arm that is controlled by an agent with access to a high-resolution video of the robot. The agent's observation intertwines information that the agent controls with exogenous content such as lighting conditions or videos in the background. We define the *control-endogenous latent state* as the parsimonious representation, which includes only information that either can be controlled by the agent (such as an object on the table that can be manipulated by the arm) or affects the agent's control (e.g., an obstacle blocking the robot arm's motion). Discovering this representation while ignoring irrelevant information offers the promise of vastly improved planning, exploration, and interpretability.

A key challenge to discovering these representations in real world applications, is that only a continual stream of observations along with the agent's actions are available. For example, consider discovering the control-endogenous latent state from a video of a robot arm along with the sequence of actions taken to control the robot. Efroni et al. (2022c) demonstrated an algorithm for discovering control-endogenous latent state that uses open-loop planning with the agent reset to a fixed start state at the beginning of each episode, which is impractical in most real domains. Prior works (Efroni et al., 2022a; Wang et al., 2022b) also discovered the control-endogenous latent state by assuming access to a given factorized encoder, which is often not available in practice.

In this work, we consider a setting where the agent acts in a world consisting of complex observations with relevant and irrelevant information. We consider environments where observations are generated from states that can be decoupled into *control-endogenous* and *exogenous* components. Our definition of exogenous

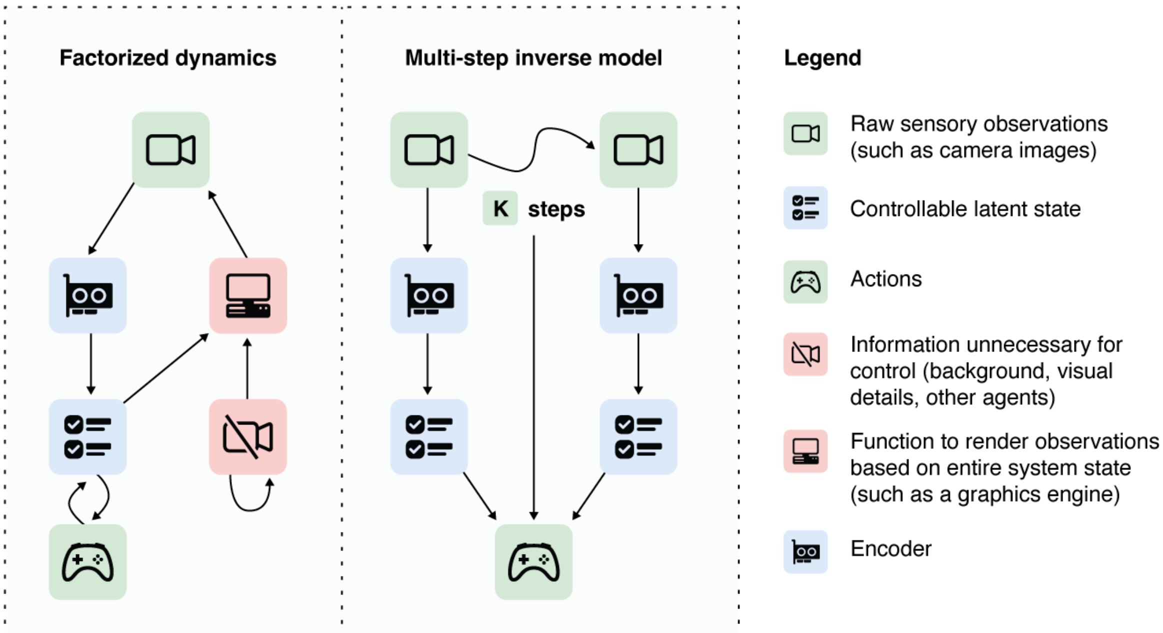

Figure 1: The control-endogenous latent state (left) can be discovered using `AC-State` (right).

strictly refers to aspects of the world that can never interact with the agent, which may include things like the detailed visual textures of objects or background processes that the agent cannot interact with. It is vital to discard irrelevant parts of the observation space. Our proposed algorithm predicts the first action that needs to be taken from any current observation to reach any future state in a certain number of steps in the future. We establish some theoretical results showing that this captures a complete and minimal control-endogenous state. This has many appealing properties in that we never need to learn a sequence model or a generative model. We only need to use supervised learning to predict actions. Prior works have considered partially observable settings to summarize agent-states (Littman et al., 1995; Kaelbling et al., 1998), or quantify agent states based on predictions of rewards (Littman et al., 2001). We emphasize here the salience of information for controlling the agent in fully-observed settings.

How can we discover this control-endogenous latent state from raw observations and actions? We introduce the Agent Control-Endogenous (`AC-State`) algorithm, which provably guarantees discovery of the control-endogenous latent state by excluding all aspects of the observations that are unnecessary for control. `AC-State` learns an encoder f that maps a given observation x to the corresponding control-endogenous latent state $f(x)$. This is accomplished by optimizing a novel objective using an expressive model class such as deep neural networks (LeCun et al., 2015). The proposed algorithm uses a multi-step inverse model as its objective, along with a bottleneck on the capacity of the learned representation.

In our experiments, we demonstrate the discovery of agent control-endogenous latent states that are (nearly) identical to their ground-truth values, while only using access to the raw observations and actions. We measure the exactness of the correspondence between the ground truth latent state (not used for training) and the learned latent state. The `AC-State` approach requires *no* reward signal, and many experiments require a few thousand interactions with the environment for initial learning along with *zero* additional examples to discover an appropriate policy. At the same time, successful discovery of the latent state implies the ability to localize, explore, and plan to reach goal states from a single shared representation.

2 Exogenous Block MDP Setting

We consider the Exogenous Block Markov Decision Process (Ex-BMDP) setting to model systems with control-endogenous and exogenous subsystems, as formulated in Efroni et al. (2022c). This definition assumes that exogenous noise has no interaction with the control-endogenous state, which is stricter than the exogenous MDP definition introduced by Dietterich et al. (2018).

Notation and Assumptions: A BMDP consists of a set of observations, $\mathcal{X}$; a set of latent states, $\mathcal{Z}$; a finite set of actions, $\mathcal{A}$ with cardinality A; a transition function, $T : \mathcal{Z} \times \mathcal{A} \to \Delta(\mathcal{Z})$; an emission function $q : \mathcal{Z} \to \Delta(\mathcal{X})$; a reward function $R : \mathcal{X} \times \mathcal{A} \to [0, 1]$; and a start state distribution $\mu_0 \in \Delta(\mathcal{Z})$. We do not consider the episodic setting, but only assume access to a single trajectory. The agent interacts with the environment, generating an observation and action sequence, $(z_1, x_1, a_1, z_2, x_2, a_2, \cdots)$ where $z_1 \sim \mu(\cdot)$. The latent dynamics follow $z_{t+1} \sim T(z' \mid z_t, a_t)$ and observations are generated from the latent state at the same time step: $x_t \sim q(\cdot \mid z_t)$. The agent does not observe the latent states $(z_1, z_2, \cdots)$, instead it receives only the observations $(x_1, x_2, \cdots)$. The *block assumption* holds if the support of the emission distributions of any two latent states are disjoint, $\text{supp}(q(\cdot \mid z_1)) \cap \text{supp}(q(\cdot|z_2)) = \emptyset$ when $z_1 \neq z_2$., where $\text{supp}(q(\cdot \mid z)) = \{x \in \mathcal{X} \mid q(x \mid z) > 0\}$ for any latent state z. Lastly, the agent chooses actions using a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, so that $a_t \sim \pi(\cdot|x_t)$.

We now define the model we consider in this work, which we refer as *deterministic Ex-BMDP*:

Definition 1 (Deterministic Ex-BMDP)**.** *A deterministic Ex-BMDP is a BMDP such that the latent state can be decoupled into two parts $z = (s, e)$ where $s \in \mathcal{S}$ is the control-endogenous state and $e \in \Xi$ is the exogenous state. We will further assume that the control-endogenous dynamics $T(s' \mid s, a)$ are deterministic. The exogenous dynamics $T_e(e' \mid e)$ may be stochastic. For $z, z' \in \mathcal{Z}, a \in \mathcal{A}$ the transition function is decoupled $T(z' \mid z, a) = T(s' \mid s, a)T_e(e' \mid e)$.*

The above definition implies that there exist mappings $f_\star : \mathcal{X} \to [S]$ and $f_{\star,e} : \mathcal{X} \to [E]$ from observations to the corresponding control-endogenous and exogenous latent states. The cardinality of the exogenous latent state E may be arbitrarily large. Furthermore, we assume that the diameter of the control-endogenous part of the state space is bounded. In other words, there is an optimal policy to reach any state from any other state in a finite number of steps:

Assumption 2 (Bounded Diameter of Control-Endogenous State Space)**.** *The length of the shortest path between any $z_1 \in \mathcal{S}$ to any $z_2 \in \mathcal{S}$ is bounded by D.*

We now describe a structural result of the Ex-BMDP model, proved in Efroni et al. (2022c). We say that π is an *endogenous policy* if it is not a function of the exogenous noise. Formally, for any x_1 and x_2, if $f_\star(x_1) = f_\star(x_2)$ then $\pi(\cdot \mid x_1) = \pi(\cdot \mid f_\star(x_2))$.

Let $\mathbb{P}_\pi(s' \mid s, t)$ be the probability to observe the control-endogenous latent state $s = f_\star(x')$, t time steps after observing $s' = f_\star(x)$ and following policy π. The following result shows that, when executing an endogenous policy, the future t time step distribution of the observation process conditioning on any x has a decoupling property. Using this decoupling property we later prove that the control-endogenous state partition is sufficient to minimize the loss of the `AC-State` objective.

Proposition 3 (Factorization Property under an Endogenous Policy, (Efroni et al., 2022c), Proposition 3)**.** *Assume that $x \sim \mu(x)$ where μ is some distribution over the observation space and that π is an endogenous policy. Then, for any $t \geq 1$ it holds that:*

$$\mathbb{P}_\pi(x' \mid x, t) = q(x' \mid f_\star(x'), f_{\star,e}(x'))\mathbb{P}_\pi(f_\star(x') \mid f_\star(x), t)\mathbb{P}(f_{\star,e}(x') \mid f_{\star,e}(x), t).$$

Additionally, we assume that the initial distribution at time step $t = 0$ is decoupled $\mu_0(s, e) = \mu_0(e)\mu_0(e)$.

Justification of Assumptions: An intuitive justification for the concept of a control-endogenous latent state comes from the design of video games. A video game typically consists of a factorized game engine (which receives player input and updates a game state) and a graphics engine. Modern video game programs are often dozens of gigabytes, with nearly all of the space being used to store high-resolution textures, audio files, and other assets. A full generative model $T(x_{t+1}|x_t, a_t)$ would need to use most of its capacity to model these high-resolution assets. On the other hand, the core gameplay engine necessary for characterizing the control-endogenous latent dynamics $T(s_{t+1} \mid s_t, a_t)$ takes up only a small fraction of the overall game program's size and thus may be much easier to model[1]. To give a concrete example, when the 1993 game *Link's Awakening* was remade in 2019 with similar gameplay but updated graphics, its file size increased from 8 MB to 5.8 GB, a 725x increase in size (Michel, 2022). Successful reinforcement learning projects for modern video games such as AlphaStar (Starcraft 2) and OpenAI Five (Dota 2) take advantage of this gap by learning on top of internal game states rather than raw visual observations (Vinyals et al., 2019; Berner et al., 2019). In the real world, there is no internal game state that we can extract. The control-endogenous latent state must be learned from experience.

Additionally, we make the further assumption that the data is collected under an endogenous policy. In this paper, we consider two kinds of policies: one is a random rollout policy, which is trivially an endogenous policy. The other type of policy we consider is achieved by planning to reach self-defined goals. This is achieved using a tabular-MDP constructed from counts of observed (s, a, s') tuples, where s is the discretized output of our learned encoder. The endogenous policy assumption in our theory may not always hold in practice, as a learned encoder may have errors that cause it to depend on exogenous noise.

We have also assumed that the control-endogenous state space has a finite state, a bounded diameter, and deterministic dynamics. The assumption of a finite state rules out continuous control problems and also makes problems with combinatorial structures (such as an arm with many joints or a robot interacting with multiple objects) infeasible. The requirement of a bounded diameter rules out environments that are unsafe or allow for irreversible changes, such as a robot becoming permanently stuck or damaged. The assumption of deterministic control-endogenous dynamics is not strictly required to use `AC-State`, but is required for a part of the proof that requires constructing the set of states that are reachable in a certain number of steps.

Goal: Our goal is reward-free and task-independent latent state discovery. Successful discovery of the control-endogenous latent state entails only learning a model for how to encode s such that $T(s_{t+1} \mid s_t, a_t)$, while not learning anything about $T(e_{t+1} \mid e_t)$ or how to encode e. In particular, we seek to show that the control-endogenous latent state is identifiable and can be recovered exactly in the asymptotic regime.

3 Prior Approaches

Deep learning architectures can be optimized for a wide range of differentiable objective functions. Our key question is: what is an objective for provably learning a control-endogenous latent state that is compatible with deep learning? At issue is finding parsimonious representations that are sufficient for control of a dynamical system given observations from rich sensors (such as high-resolution videos) while discarding irrelevant details. Approaches such as optimal state estimation (Durrant-Whyte & Bailey, 2006), system identification (Ljung, 1998), and simultaneous localization and mapping (Cadena et al., 2016; Dissanayake et al., 2001) achieve parsimonious state estimation for control, yet require more domain expertise and design than is desirable. Previous learning-based approaches failed to capture the full control-endogenous latent state or exclude all irrelevant information (Efroni et al., 2022c). Reinforcement learning approaches that capture the latent state from rich sensors by employing autoencoders (Goodfellow et al., 2016) or contrastive learning often capture noise components[2].

Learning latent states for interactive environments is a mature research area with prolific contributions. We discuss a few of the most important lines of research and how they fail to achieve guaranteed discovery of the control-endogenous latent state. We categorize these contributions into broad areas based on what they predict: rewards (or value functions), future latent states, observations, the relationship between observations, and/or actions. Table 1 provides a summarized comparison of the properties of `AC-State` and prior works. In this table, algorithms that must learn the full exogenous state before learning how to discard it are marked as having an exogenous invariant state but not exogenous invariant learning.

[1] The control-endogenous latent state and their dynamics is generally even more compact than the game state, which itself may have exogenous noise such as unused or redundant variables.

[2] Consider a divided freeway, where cars travel on opposing sides of the lane. For this situation, autoencoders or contrastive learning objectives (commonly referred to as "self-supervised learning") produce distinct latent states for every unique configuration of cars on the other side of the lane divider. For example, an autoencoder or generative model would learn to predict the full configuration and visual details of all the cars, even those that could not interact with the agent.

Limitations of Predicting Rewards: Reward-based bisimulation (Zhang et al., 2021b) can filter irrelevant information from the latent state but is dependent on access to a reward signal. Deep reinforcement learning based on reward optimization (Mnih et al.) often struggles with sample complexity when the reward is sparse and fails completely when the reward is absent.

Limitations of Predicting Latent States: Approaches that involve predicting future latent states from past latent states have achieved good performance, but there is no theoretical guarantee that the latent state captures the full control-endogenous latent state (Guo et al., 2022; Schwarzer et al., 2021b; Ye et al., 2021; Pathak et al., 2017). Deep bisimulation approaches learn state representations for control tasks with agnosticism toward task-irrelevant details (Zhang et al., 2021a). Prior approaches have involved a model that predicts latent states such that the pre-trained representations can solve a task (Schwarzer et al., 2021a), yet this approach is not task-agnostic, and it is not guaranteed to recover the control-endogenous state. An auto-encoder trained with reconstruction loss or a dynamics model (Lange et al., 2012; Wahlström et al., 2015; Watter et al., 2015) learns low-dimensional state representations to capture information relevant for solving a task. Although learning latent state representations has been shown to be useful for solving tasks, there is no guarantee that such methods can fully recover the underlying control-endogenous latent state.

Limitations of Predicting Observations: Other works have also used generative models or autoencoders to predict future observations for learning latent representations, mostly for purposes of exploration. The idea of predicting observations is often referred to as "intrinsic motivation" to guide the agent towards exploring unseen regions of state space (Oudeyer & Kaplan, 2007). Other works use autoencoders to estimate future observations in the feature space for exploration (Stadie et al., 2015), though such models can also fail in the presence of exogenous observations. Dynamics models learn to predict distributions over future observations, but often for the purpose of planning a sequence of actions instead of recovering latent states or for purposes of exploration. Some approaches that learn generative models in the observation space attempt to learn a further decomposition of the dynamics into into control-endogenous and exogenous latent states. While this can result in the correct latent state, it still requires learning a full generative model over the entire latent state, not just the control-endogenous latent state (Wang et al., 2022a;b).

Limitations of Predicting Relationships between Observations: By learning to predict relations between two consecutive observations, prior works have attempted to learn latent state representations, both theoretically (Misra et al., 2020) and empirically (Mazoure et al., 2020). For example, by exploiting mutual information based objectives (information gain based on current states and actions with future states) (Mazoure et al., 2020; Song et al., 2012), previous works have attempted to learn control-endogenous states in the presence of exogenous noise. However, unlike `AC-State`, they learn latent states dependent on exogenous noise, even though the learned representation can be useful for solving complex tasks (Mazoure et al., 2020). Theoretically, Misra et al. (2020) uses a contrastive loss based objective to provably learn latent state representations that can be useful for hard exploration tasks. However, contrastive loss based representations can still be prone to exogenous noise (Efroni et al., 2022c), whereas `AC-State` exploits an exogenous free rollout policy with a multi-step inverse dynamics model to provably and experimentally recover the full control-endogenous latent state.

Limitations of Predicting Actions: `AC-State` aims at recovering the control-endogenous latent states by training a multi-step inverse dynamics model in the presence of exogenous noise. While prior works have explored similar objectives, either for exploration or for learning state representations, they are unable to recover latent states with perfect accuracy in the presence of exogenous noise (Efroni et al., 2022c). In Appendix B, we provide a concrete counter-example demonstrating why a one-step inverse model might fail. Approaches based on action prediction with one-step inverse models (Pathak et al., 2017) are widely used in practice (Baker et al., 2022; Badia et al., 2020). However, these inverse models can fail to capture the full control-endogenous latent dynamics (Efroni et al., 2022c; Hutter & Hansen, 2022), while combining them with an autoencoder (Bharadhwaj et al., 2022) inherits the weaknesses of that approach.

Limitations of Empowerment-based objectives : Empowerment-based objectives focus on the idea that an agent should try to seek out states where it is empowered by having the greatest number of possible states that it can easily reach (Klyubin et al., 2005). For example, in a maze with two rooms, the most

Algorithms	PPE	OSSR	DBC	CDL	Denoised-MDP	1-Step Inverse	AC-State (Ours)
Exogenous Invariant State	✓	✓	✓	✓	✓	✓	✓
Exogenous Invariant Learning	✓	✓	✗	✗	✗	✓	✓
Flexible Encoder	✓	✗	✓	✗	✓	✓	✓
YOLO (No Resets) Setting	✗	✓	✓	✓	✓	✓	✓
Reward Free	✓	✓	✗	✓	✓	✓	✓
Control-Endogenous Rep.	✓	✓	✗	✓	✓	✗	✓

Table 1: **An Overview of the Properties** of prior works on representation learning in RL is shown, with a particular emphasis on robustness to exogenous information. The comparison to `AC-State` aims to be as generous as possible to the baselines. ✗ is used to indicate a known counterexample for a given property.

empowered state is the doorway, since it makes it easy to reach either of the rooms. Concrete instantiations of the empowerment objective may involve training models to predict the distribution of actions from observations (either single-step or multi-step inverse models) (Mohamed & Rezende, 2015; Yu et al., 2019), but they lack the information bottleneck term and the requirement of an exogenous-independent rollout policy. The analysis and theory in this work focus on action prediction as a particular method for measuring empowerment rather than as a way of guaranteeing the discovery of a minimal control-endogenous latent state and ignoring exogenous noise.

Limitations of other approaches using Multi-Step Inverse Models: Here we discuss several recent works that study the reinforcement learning problem in the presence of exogenous and irrelevant information. Efroni et al. (2022c) formulated the Ex-BMDP model and designed a provably efficient algorithm that learns the state representation. However, their algorithm succeeds only in the episodic setting, when the initial control-endogenous state is initialized deterministically. This strict assumption makes their algorithm impractical in many cases of interest. Here we removed the deterministic assumption of the initial latent state. Indeed, as our theory suggests, `AC-State` can be applied in the "you-only-live-once" (YOLO) setting when an agent has access to a single trajectory. In Efroni et al. (2022b) and OSSR (Efroni et al., 2022a) the authors designed provable RL algorithms that efficiently learn in the presence of exogenous noise under different assumptions about the underlying dynamics. These works, however, focused on statistical aspects of the problem; how to scale these approaches to complex environments and combine function approximations is currently unknown and seems challenging.

Significance of our work: `AC-State` uses a simple multi-step inverse model along with a bottleneck on the representations to learn latent states, which has not been exploited by prior works using dynamics models. Unlike the aforementioned works, here we focus on the representation learning problem. `AC-State` predicts actions based on future observations, and we show that a multi-step inverse dynamics model predicting actions can fully recover the control-endogenous latent states with no dependence on the exogenous noise. We emphasize that in the presence of exogenous noise, methods based on predictive future observations are prone to predicting both the control-endogenous and exogenous parts of the state space and do not have guarantees on recovering the latent structure. `AC-State` is the first practical and theoretically grounded approach that learns the control-endogenous representation with complex function approximators such as deep neural networks.

4 AC-State: Discovering Agent Control-Endogenous Latent States

The control-endogenous latent state should preserve information about the interaction between states and actions while discarding irrelevant details. The proposed objective (Agent Control-Endogenous State or *AC-State*) accomplishes this by generalizing one-step inverse dynamics (Pathak et al., 2017; Efroni et al., 2022c; Hutter & Hansen, 2022; Badia et al., 2020) to multiple steps with an information bottleneck (Figure 1). This multi-step inverse model predicts the first action taken to reach a control-endogenous latent state k steps in the future:

$$\mathcal{L}_{\texttt{AC-State}}(f, x, a, t, k) = -\log(\mathbb{P}(a_t | f(x_t), f(x_{t+k}); k)) \tag{1}$$

The multi-step inverse objective $\mathcal{L}_{\texttt{AC-State}}$ predicts the action a_t from the observation just before the action a_t is taken and from an observation collected k steps in the future. The action step t is sampled uniformly over all of the $\mathcal{T}$ steps in which the action was taken randomly. The prediction horizon k is sampled uniformly from 1 to a maximum horizon K_t, which may be a fixed hyperparameter or may depend on the number of steps left in the planning horizon at step t. The multi-step inverse model predicts the distribution over the first action to reach a state k steps in the future from a given current state. Generally, this action cannot be predicted perfectly, but it still carries information about the relationship between the current state and a state in the (potentially distant) future.

We optimize the parameters of an encoder model $f : \mathbb{R}^n \to \{1, \ldots, \text{Range}(f)\}$ that maps from an n-dimensional continuous observation to a finite latent state with an output range of size Range(f), an integer number. $\mathcal{F}$ is a set that represents all mappings achievable by the model. We wish to minimize Objective 1 while using an encoder with the smallest latent state as its output. Conceptually, we can define this as finding a set of optimal solutions. The control-endogenous latent state is guaranteed to be one of the optimal solutions for the multi-step inverse objective $\mathcal{L}_{\texttt{AC-State}}$, while other solutions may fail to remove irrelevant information. In the robot example from earlier, a minimal latent state ignores background distractions, irrelevant visual details, and objects that are unrelated to the agent (Figure 2). The control-endogenous latent state is uniquely achievable by finding the lowest capacity f in the set of optimal solutions to `AC-State` objective:

Figure 2: `AC-State` discovers a control-endogenous latent state in a visual robotic setting with temporally correlated distractors: a TV, flashing lights, and drinking bird toys (left). We visualize the learned latent state by training a decoder to reconstruct the observation (right). This shows that `AC-State` learns to discover the robotic arm's position while ignoring the background distractions.

$$G = \arg\min_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{t \sim \mathcal{T}} \mathop{\mathbb{E}}_{k \sim U(1, K_t)} \mathcal{L}_{\texttt{AC-State}}(f, x, a, t, k) \tag{2}$$

$$\widehat{f} \in \arg\min_{f \in G} \text{Range}(f). \tag{3}$$

As a concrete strategy to achieve a minimal latent state, we combine two mechanisms that are widely used in the deep learning literature to restrict the information capacity in f, which introduces an additional loss term $\mathcal{L}_{\text{Bottleneck}}(f, x_t)$. We first pass the hidden state at the end of the network through a gaussian variational information bottleneck (Alemi et al., 2017), which reduces the mutual information between x and the representation. We then apply vector quantization (Van Den Oord et al., 2017), which yields a discrete latent state. While either of these could be used on their own, we found that adding the gaussian mutual information objective eased the discovery of parsimonious discrete representations.

In addition to optimizing the objective and restricting capacity, the actions taken by the agent are also important for the success of `AC-State` , in that they must achieve high coverage of the control-endogenous state space and not depend on the exogenous noise. This is satisfied by a random policy or a policy that depends on $\widehat{f}(x_t)$ and achieves high coverage. The `AC-State` objective enjoys provable asymptotic success (see supplementary materials) in discovering the control-endogenous latent state.

Intuitively, the `AC-State` objective encourages the latent state to keep information about the long-term effect of actions, which requires storing all information about how the actions affect the world. At the same time,

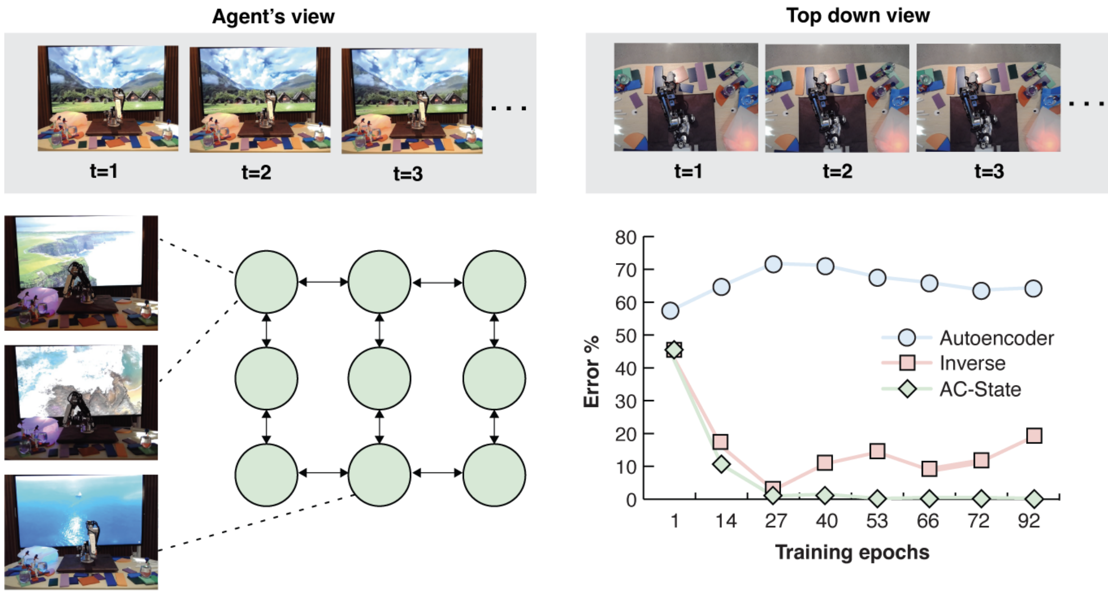

Figure 3: A real robotic arm moves between nine different positions (left). The quality of control-endogenous latent state dynamics learned by `AC-State` is better than one-step inverse models and autoencoders (bottom right).

the `AC-State` objective never requires predicting information about the observations themselves, so it places no value on representing aspects of the world that are unrelated to the agent.

AC-State with Random Policy: The simpler version of AC-State uses a random policy to collect data and fit the encoder (Algorithm 1). While this is simple to implement, a uniform random policy is always exogenous-free by construction and asymptotically achieves high coverage. At the same time, it may not achieve sufficient coverage with reasonable sample efficiency for some hard exploration tasks. If a single bad action can cause an agent to lose many steps of progress, the sample complexity needed to achieve high coverage can scale exponentially in the horizon.

AC-State with Planning Policy: A more involved algorithm is able to discover latent states in problems where exploration is difficult (such as in an environment where specific actions need to be taken in sequence to make progress). This can be accomplished by using a planning policy instead of a purely random policy. While any type of policy could be used in principle, we experimented with planning in a simple count-based tabular-MDP. In this model, a monte-carlo breadth-first search (suitable for stochastic dynamics) is used to select rarely seen reachable states as goal states. Then Dijkstra's algorithm is used to construct closed-loop plans to reach the desired goal state. Only a single random action is taken on the first step of each goal-seeking trajectory, and it is this random action that is predicted with the `AC-State` objective. While the `AC-State` theory assumes deterministic dynamics, the tabular-MDP constructed using learned latent states often has stochastic dynamics due to errors in the learned encoder.

5 Theoretical Analysis

We present an asymptotic analysis of `AC-State` showing it recovers $f_\star$, the control-endogenous state representation. The mathematical model we consider is the deterministic Ex-BMDP. There, the transition model of the latent state decomposes into a control-endogenous latent state, which evolves deterministically, along with a noise term–the agent-irrelevant portion of the state. The noise term may be an arbitrary temporally correlated stochastic process. If the reward does not depend on this noise, any optimal policy may be expressed in terms of this control-endogenous latent state. In this sense, the recovered control-endogenous latent state is sufficient for achieving optimal behavior.

The proof involves two steps. In the first step, we show that a Bayes-optimal solution to the multi-step inverse model objective can be achieved without dependence on exogenous noise. In the second step, we show that the minimal representation minimizing the multi-step inverse model objective is the control-endogenous state.

5.1 Bayes Optimal Solution does not depend on Exogenous Noise

Consider the generative process in which x is sampled from a distribution μ, the agent executes a policy π for t time steps and samples x'. Denote by $\mathbb{P}_{\pi,\mu}(x, x', t)$ as the joint probability, and by $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ as the probability that under this generative process the action upon observing x is a. The following result, which builds on Proposition 3, shows that the optimal bayes solution $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ is equal to $\mathbb{P}_{\pi,\mu}(a \mid f_\star(x), f_\star(x'), t)$ for $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$, where $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ is the probability to sample x .

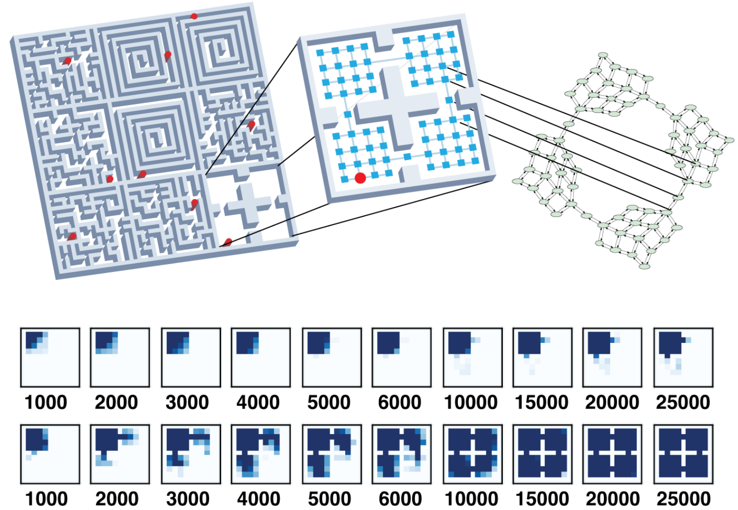

Figure 4: We study a multi-agent world where each of the nine mazes has a separately controlled agent. Training `AC-State` with the actions of a specific agent discovers its control-endogenous latent state while discarding information about the other mazes (top). In a version of this environment where a fixed third of the actions cause the agent's position to reset to the top-left corner, a random policy fails to explore, whereas planned exploration using `AC-State` reaches all parts of the maze (bottom).

Proposition 4. *Assume that π is an endogenous policy. Let $x \sim \mu$ for some distribution μ. Then, the Bayes' optimal predictor of the action-prediction model is piece-wise constant with respect to the control-endogenous partition: for all $a \in \mathcal{A}$, $t > 0$ and $x, x' \in \mathcal{X}$ such that $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ it holds that*

$$\mathbb{P}_{\pi,\mu}(a \mid x, x', t) = \mathbb{P}_{\pi,\mu}(a \mid f_\star(x), f_\star(x'), t).$$

We comment that the condition $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ is necessary since ,otherwise, the conditional probability $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ is well not defined.

The proof for this proposition is a straightforward application of bayes theorem, is given in the appendix, and was previously presented in Efroni et al. (2022c;a).

5.2 Discovery of Full Control-Endogenous Latent State

Proposition 9 from the previous section shows that the multi-step action-prediction model is piecewise constant with respect to the partition induced by the control-endogenous states $f_\star : \mathcal{X} \to [S]$. In this section, we assume that the executed policy is an endogenous policy that induces sufficient exploration. With this, we prove that there is no coarser partition of the observation space such that the set of inverse models is piecewise constant with respect to it.

We assume that the Markov chain induced on the control-endogenous state space by executing the policy $\pi_{\mathcal{D}}$ by which *AC-State* collects the data has a stationary distribution $\mu_{\mathcal{D}}$ such that $\mu_{\mathcal{D}}(s, a) > 0$ and $\pi_{\mathcal{D}}(a \mid s) \geq \pi_{\min}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. We consider the stochastic process in which an observation is sampled from a distribution μ such that $\mu(s) = \mu_{\mathcal{D}}(s)$ for all s. Then, the agent executes the policy $\pi_{\mathcal{D}}$ for t time steps. For brevity, we denote the probability measure induced by this process as $\mathbb{P}_{\mathcal{D}}$. A more detailed justification for this assumption is given in the appendix.

We begin by defining several useful notions. We denote the set of reachable control-endogenous states from s in h time steps as $\mathcal{R}_h(s)$.

Definition 5 (Reachable Control-endogenous States)**.** *Let the set of reachable control-endogenous states from $s \in \mathcal{S}$ in $h > 0$ time steps be $\mathcal{R}(s, h) = \{s' \mid \max_\pi \mathbb{P}_\pi(s' \mid s, h) = 1\}$.*

Observe that every reachable state from s in h time steps satisfies that $\max_\pi \mathbb{P}_{\pi,\mu}(s' \mid s_0 = s, h) = 1$ due to the deterministic assumption of the control-endogenous dynamics.

Next, we define a notion of *consistent partition* with respect to a set of function values. Intuitively, a partition of space $\mathcal{X}$ is consistent with a set of function values if the function is piece-wise constant on that partition.

Definition 6 (Consistent Partition with respect to $\mathcal{G}$)**.** *Consider a set $\mathcal{G} = \{g(a, y, y')\}_{y,y' \in \mathcal{Y}, a \in \mathcal{A}}$ where $g : \mathcal{A} \times \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. We say that $f : \mathcal{Y} \to [N]$ is a* consistent partition *with respect to $\mathcal{G}$ if for all $y, y'_1, y'_2 \in \mathcal{Y}$, $f(y'_1) = f(y'_2)$ implies that $g(a, y, y'_1) = g(a, y, y'_2)$ for all $a \in \mathcal{A}$.*

Observe that Proposition 9 shows that the partition of $\mathcal{X}$ according to $f_\star$ is consistent with respect to $\{\mathbb{P}_\mathcal{D}(a \mid x, x', h) \mid x, x' \in \mathcal{X}, h \in [H] \text{ s.t. } \mathbb{P}_\mathcal{D}(x, x', h) > 0\}$, since, by Proposition 9, $\mathbb{P}_\mathcal{D}(a \mid x, x', h) = \mathbb{P}_\mathcal{D}(a \mid f_\star(x), f_\star(x'), h)$.

Towards establishing that the coarsest abstraction according to the `AC-State` objective is $f_\star$ we make the following definition.

Definition 7 (The Generalized Inverse Dynamics Set $\mathrm{AC}(s, h)$)**.** *Let $s \in \mathcal{S}, h \in \mathbb{N}$. We denote by* $\mathrm{AC}(s, h)$ *as the set of multi-step inverse models accessible from s in h time steps. Formally,*

$$\mathrm{AC}(s, h) = \{\mathbb{P}_\mathcal{D}(a \mid s', s'', h') : s' \in \mathcal{R}(s, h - h'), s'' \in \mathcal{R}(s', h'), a \in \mathcal{A}, h' \in [h]\}. \tag{4}$$

Observe that in equation equation 5 the inverse function $\mathbb{P}_\mathcal{D}(a \mid s', s'', h')$ is always well defined since $\mathbb{P}_\mathcal{D}(s', s'', h') > 0$. It holds that $\mathbb{P}_\mathcal{D}(s', s'', h') = \mathbb{P}_\mathcal{D}(s')\mathbb{P}_\mathcal{D}(s'' \mid s', h') > 0$, since $\mathbb{P}_\mathcal{D}(s') > 0$ and $\mathbb{P}_\mathcal{D}(s'' \mid s', h') > 0$. The inequality $\mathbb{P}_\mathcal{D}(s') > 0$ holds by the assumption that the stationary distribution when following $\mathcal{U}$ has positive support on all control-endogenous states (Assumption 10). The inequality $\mathbb{P}_\mathcal{D}(s'' \mid s', h') > 0$ holds since, by definition $s'' \in \mathcal{R}(s, h')$ is reachable from s' in h' time steps; hence, $\mathbb{P}_\mathcal{D}(s'' \mid s', h') > 0$ by the fact that Assumption 10 implies that the policy $\pi_\mathcal{D}$ induces sufficient exploration.

Theorem 8 ($f_\star$ is the coarsest partition consistent with respect to *AC-State* objective)**.** *Given that Assumption 2 (bounded diameter) and Assumption 10 (coverage) hold, it follows that there is no coarser partition than $f_\star$, which is consistent with* $\mathrm{AC}(s, D)$ *for any $s \in \mathcal{S}$.*

The proof of this theorem, which uses induction on h, is provided in the appendix.

6 Experiments

Our experiments explore three domains and demonstrate the unique capabilities of `AC-State`: (i) `AC-State` learns the control-endogenous latent state of a real robot from a high-resolution video of the robot with rich temporal background structure (a TV playing a video, flashing lights, dipping birds, and even people) *occluding the information in the observations*; (ii) `AC-State` learns about the controlled agent while ignoring other random agents in a multiple maze environment where there are other functionally identical agents. We demonstrate that `AC-State` solves a hard maze exploration problem. (iii) `AC-State` learns a control-endogenous latent state in a house navigation environment where the observations are high-resolution images and the camera's vertical position randomly oscillates, showing that `AC-State` is invariant to exogenous viewpoint noise that radically changes the observation.

6.1 Robotic Arm Experiment

We found that `AC-State` discovers the control-endogenous latent state of a real robot arm while ignoring distractions. We collected 6 hours of data (14,000 samples) from the robot arm (Annin) by taking four high-level actions (move left, move right, move up, and move down). A picture of the robot was taken after each completed action. Details of our experiment setup are provided in Appendix A.2. The robot was placed among many distractions, such as a television, flashing, color-changing lights, and people moving in the background. Some of these distractions (especially the TV) had strong temporal correlations between adjacent time steps, as is often the case in real-life situations. We discovered the control-endogenous latent state (Figure 3), which is the ground truth position of the robot (not used during training). We qualitatively evaluated `AC-State` by training a convolutional neural network to reconstruct x from $f(x)$ with square loss

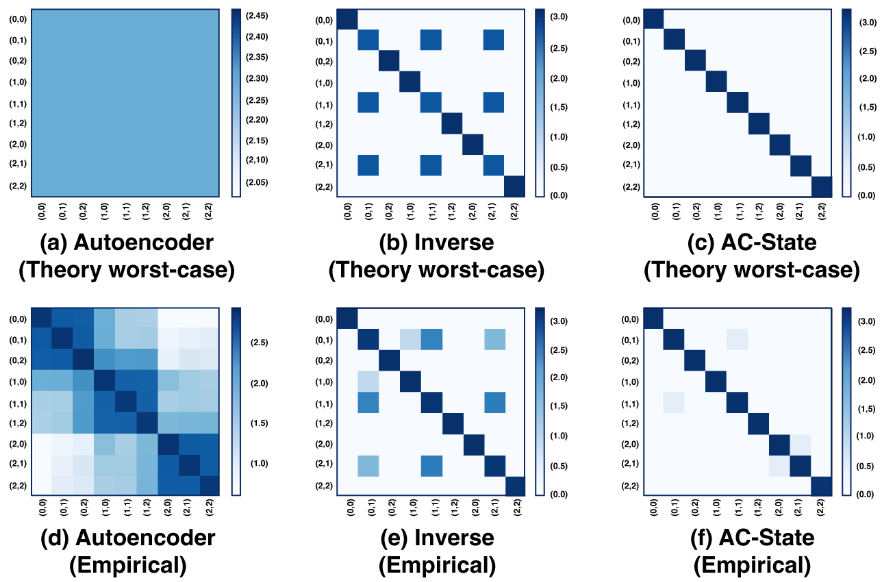

Figure 5: In co-occurrence histograms measuring performance in the robotic arm environment, autoencoders fail (left), one-step inverse models fall prey to the same counterexample in theory and in experiment (center), and `AC-State` discovers a perfect control-endogenous latent state (diagonal histogram, right).

as the error metric. We found that the robot arm's position was correctly reconstructed, while the distracting TV and color-changing lights appeared completely blank, as expected (Figure 2). As discussed previously, an auto-encoder trained end-to-end with x as input captures both the control-endogenous latent state and distractor noise. The theoretical counter-examples for autoencoders and one-step inverse models also closely match the errors that we observe experimentally (Figure 5).

6.2 Mazes with Exogenous Agents and Reset Actions

In our experiments, `AC-State` removes exogenous noise in environments where that noise has a rich and complex structure. In particular, we studied a multi-agent system in which a single agent is control-endogenous and the other agents follow their own independent policies. In an environment with 9 agents and each agent having c control-endogenous states, the overall size of the observation space is c^9. With 3,000 training samples, `AC-State` is able to nearly perfectly discover the agent's control-endogenous latent state, while fully ignoring the state of the 8 uncontrollable exogenous agents, with all of the agents controlled by a random policy (Figure 4).

The control-endogenous latent state is useful when it allows for exponentially more efficient exploration than is achievable under a random policy. To exhibit this, we modified the maze problem by giving the agent additional actions that reset to a fixed initial starting position. When a third of all actions cause resets, the probability of acting randomly for N steps without resetting is $(2/3)^N$. We show that a learned exploration policy using `AC-State` succeeds in full exploration and learning of the control-endogenous latent state with 25,000 samples, while a random exploration policy barely explores the first room with the same number of samples (Figure 4).

The counts of the discrete latent states are used to construct a simple tabular MDP where planning is done to reach goal states using a monte carlo version of Dijkstra's algorithm (to account for stochastic transition dynamics). The reachable goal states are sampled proportionally to $\frac{1}{count(s_i)}$, so the rarely seen states are the most likely to be selected as goals. Experiment results demonstrate that a goal-seeking policy achieves perfect coverage of the state space by using discovered latents for exploration, while a random policy fails to reach more than 25% of the state space in the presence of reset actions. We demonstrate this with heatmaps showing state visitation frequencies.

6.3 First-Person Perspective House Navigation

In order to analyze the performance of the proposed `AC-State` objective in a more realistic setting, we evaluated it on Matterport (Chang et al., 2017), a navigation environment where each observation is a high resolution image taken from a position in a real house. A 20,000 sample dataset collected from an agent moving randomly through the house is used to train `AC-State`. In addition to the high degree of visual information in the input observations, we randomly move the camera up or down at each step as a controlled source of irrelevant information (exogenous noise). `AC-State` removes the view noise from the encoded representation $f(x)$ while still capturing the true control-endogenous latent state (Figure 6), whereas other baselines capture both the control-endogenous latent state and the exogenous noise. Details on the matterport simulator are provided in Appendix A.3.

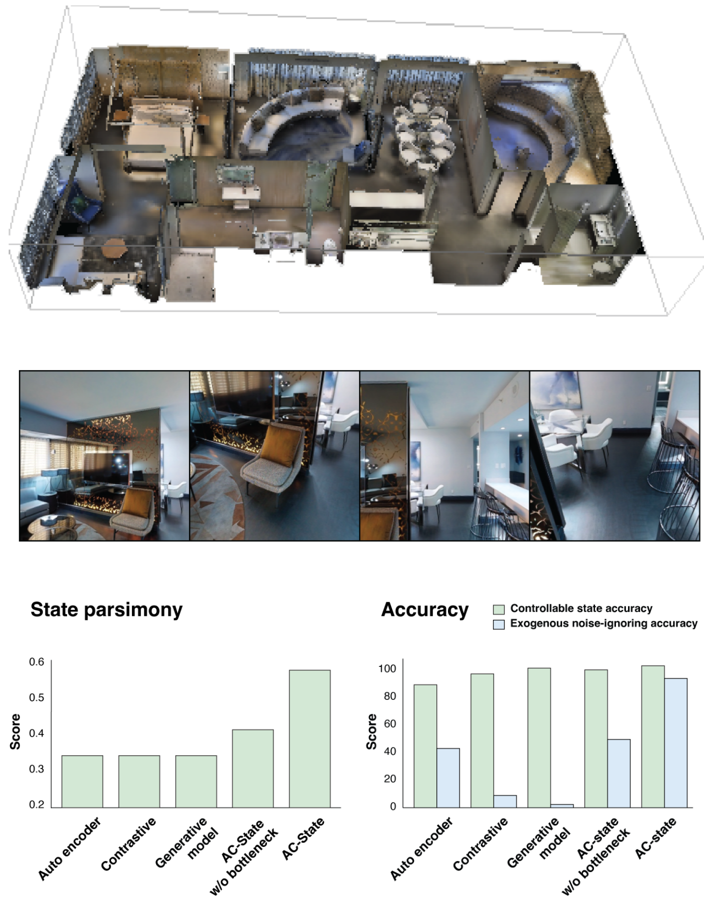

Figure 6: We evaluate `AC-State` in a house navigation environment (top), where the agent observes high-resolution images of first-person views and the vertical position of the camera is exogenous (center). `AC-State` discovers a control-endogenous latent state that is parsimonious (bottom left) and captures the position of the agent in the house while discarding the exogenous noise (bottom right). The baselines we consider capture the control-endogenous latent state but fail to discard the exogenous noise.

We present the results for this experiment in Figure 6. The *Controllable Latent State Accuracy* is the viewpoint prediction accuracy for the current state. The *Exogenous Noise-Ignoring Accuracy* reflects how much information about the exogenous noise is removed from the observation. We can see that the proposed *AC-State* model has the highest control-endogenous latent state and exogenous noise-ignoring accuracy. Thus, it outperforms the baselines we considered at capturing control-endogenous state while ignoring exogenous noise. We calculated state parsimony as $\frac{\text{Num. Ground Truth States}}{\text{Num. Discovered States}}$. Therefore, a lower state parsimony denotes a high number of discovered states, which means that the model fails at ignoring exogenous information. The proposed model has the highest state parsimony, which shows the effectiveness of the model in ignoring exogenous noise while only capturing control-endogenous latent states.

7 Conclusion

`AC-State` reliably discovers the control-endogenous latent state across multiple domains. The vast simplification of the control-endogenous latent state discovered by `AC-State` enables visualization, exact planning, and fast exploration. The field of self-supervised reinforcement learning particularly benefits from these approaches, with `AC-State` useful across a wide range of applications involving interactive agents as a self contained module to improve sample efficiency given any task specification. As the richness of sensors and the ubiquity of computing technologies (such as virtual reality, the internet of things, and self-driving cars) continues to grow, the capacity to discover agent-control endogenous latent states enables new classes of applications.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Chris Annin. Ar3 robot. Location: https://www.anninrobotics.com.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pp. 507–517. PMLR, 2020.

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. *arXiv preprint arXiv:2206.11795*, 2022.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based RL. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=DfUjyyRW90`.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

Kingma Diederik, Ba Jimmy, et al. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, pp. 273–297, 2014.

Thomas G. Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. *CoRR*, abs/1806.01584, 2018. URL `http://arxiv.org/abs/1806.01584`.

MWM Gamini Dissanayake, Paul Newman, Steve Clark, Hugh F Durrant-Whyte, and Michael Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, 2001.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *IEEE Robotics & Automation Magazine*, 13(2):99–110, 2006.

Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. *(Accepted for publication at) Conference on Learning Theory*, 2022a.

Yonathan Efroni, Sham Kakade, Akshay Krishnamurthy, and Cyril Zhang. Sparsity in partially controllable linear systems. *(Accepted for publication at) International Conference on Machine Learning*, 2022b.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022c. URL `https://openreview.net/forum?id=RQLLzMCefQu`.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT Press, 2016.

Zhaohan Daniel Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *arXiv preprint arXiv:2206.08332*, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), 2016.

Marcus Hutter and Steven Hansen. Uniqueness and complexity of inverse mdp models, 2022.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X. URL https://doi.org/10.1016/S0004-3702(98)00023-X.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All else being equal be empowered. In Mathieu S. Capcarrère, Alex A. Freitas, Peter J. Bentley, Colin G. Johnson, and Jon Timmis (eds.), *Advances in Artificial Life*, pp. 744–753, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31816-3.

Sascha Lange, Martin A. Riedmiller, and Arne Voigtländer. Autonomous reinforcement learning on raw visual input data in a real world application. In *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*, pp. 1–8. IEEE, 2012. doi: 10.1109/IJCNN.2012.6252823. URL https://doi.org/10.1109/IJCNN.2012.6252823.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Michael L. Littman, Anthony R. Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In Armand Prieditis and Stuart Russell (eds.), *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pp. 362–370. Morgan Kaufmann, 1995. doi: 10.1016/b978-1-55860-377-6.50052-9. URL https://doi.org/10.1016/b978-1-55860-377-6.50052-9.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani (eds.), *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 1555–1561. MIT Press, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

Lennart Ljung. System Identification. In *Signal Analysis and Prediction*, pp. 163–173. Springer, 1998.

Bogdan Mazoure, Remi Tachet des Combes, Thang Doan, Philip Bachman, and R. Devon Hjelm. Deep reinforcement and infomax learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/26588e932c7ccfa1df309280702fe1b5-Abstract.html.

Teddy Michel. The file size of every core legend of zelda game, 2022. URL https://www.thegamer.com/legend-of-zelda-file-size-every-game/.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. *Neural Information Processing (NIPS)*.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2125–2133, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e00406144c1e7e35240afed70f34166a-Abstract.html.

Pierre-Yves Oudeyer and Frédéric Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers Neurorobotics*, 1:6, 2007. doi: 10.3389/neuro.12.006.2007. URL https://doi.org/10.3389/neuro.12.006.2007.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which mutual-information representation learning objectives are sufficient for control? In *Advances in Neural Information Processing Systems*, volume 34, pp. 26345–26357. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/dd45045f8c68db9f54e70c67048d32e8-Paper.pdf.

Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron C. Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=uCQfPZwRaUu.

Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, Ankesh Anand, Laurent Charlin, R. Devon Hjelm, Philip Bachman, and Aaron C. Courville. Pretraining representations for data-efficient reinforcement learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12686–12699, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/69eba34671b3ef1ef38ee85caae6b2a1-Abstract.html.

Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinform.*, 13:328, 2012. doi: 10.1186/1471-2105-13-328. URL https://doi.org/10.1186/1471-2105-13-328.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *CoRR*, abs/1507.00814, 2015. URL http://arxiv.org/abs/1507.00814.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24261–24272, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Niklas Wahlström, Thomas B. Schön, and Marc Peter Deisenroth. From pixels to torques: Policy learning with deep dynamical models. *CoRR*, abs/1502.02251, 2015. URL http://arxiv.org/abs/1502.02251.

Tongzhou Wang, Simon S Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised mdps: Learning world models better than the world itself. *arXiv preprint arXiv:2206.15477*, 2022a.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction, 2022b. URL https://arxiv.org/abs/2206.13452.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin A. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2746–2754, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/a1afc58c6ca9540d057299ec3016d726-Abstract.html.

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering atari games with limited data. *Advances in Neural Information Processing Systems*, 34:25476–25488, 2021.

Tianhe Yu, Gleb Shevchuk, Dorsa Sadigh, and Chelsea Finn. Unsupervised visuomotor control through distributional planning networks. *Robotics: Science and Systems (RSS)*, 2019.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 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, 2021a. URL https://openreview.net/forum?id=-2FCwDKRREu.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=-2FCwDKRREu.

A Methods and Experiments

In this section, we will describe our methods and experiments for validating the proposed latent state discovery with `AC-State`. Three experiments, including both simulated and physical environments, are used to test the efficacy of our proposed algorithm. The environments are carefully chosen to demonstrate the ability of the `AC-State` agent to succeed at navigation and virtual manipulation tasks with varying degrees of difficulty; on these testbeds, algorithms with similar properties in literature fail to succeed. In what follows, we will describe the environmental setups, the function approximation scheme for the latent state, and the results that we produced.

A.1 Mazes with Exogenous Agents and Reset Actions

We consider a global 2D maze (see Fig. 4) further divided into nine 2D maze substructures (henceforth called gridworlds). Each gridworld is made up of 6×6 ground truth states, and only one of the gridworlds contains the `AC-State` agent. Every gridworld other than the one containing the true agent has an agent placed within it whose motion is governed by random actions. Our goal is to show that the proposed `AC-State` agent can "discover" the control-endogenous latent state within the global gridworld while ignoring the structural perturbations in the geometry of the other 8 gridworlds.

A.1.1 Exploring in Presence of Reset Actions

Data Collection: We collect data under a random roll-out policy while interacting with the gridworld's environment. We endow the agent with the ability to "reset" its action to a fixed starting state. The goal of this experiment is to show that, in the presence of reset actions, it is sufficiently hard for a random rollout policy to get full coverage of the mazes. To achieve sufficient coverage, we can leverage the discovered control-endogenous latent states to learn a goal seeking policy that can be incentivized to deterministically reach unseen regions of the state space. The counts of the discrete latent states are used to construct a simple tabular MDP where planning is done to reach goal states using a monte carlo version of Dijkstra's algorithm (to account for stochastic transition dynamics). Experimental results demonstrate that a goal-seeking policy achieves perfect coverage of the state space by using discovered latents for exploration, while a random policy fails to reach more than 25% of the state space in the presence of reset actions. We demonstrate this with heatmaps showing state visitation frequencies.

Experiment Details: The encoder receives observations of size $80 \times 720 \times 3$ due to the observations from 8 other exogenous agents. The agent has an action space of 4, where actions are picked randomly from a uniform policy. For the reset action setting, we use an additional 4 reset actions, and uniformly picking a reset action can reset it to a deterministic starting state. The observation is encoded using the MLP-Mixer architecture (Tolstikhin et al., 2021) with gated residual connections (Jang et al., 2017). The model is trained using the Adam optimizer (Diederik et al., 2014) with a default learning rate of 0.0001 and without weight decay. We use a 2-layer feed-forward network (FFN) with 512 hidden units for the encoder network, followed by a vector quantization (VQ-VAE) bottleneck. The use of a VQ-VAE bottleneck would discretize the representation from the multi-step inverse model by adding a codebook of discrete learnable codes. For recovering control-endogenous latents from the maze we want to control while ignoring the other exogenous mazes, we further use a MLP-Mixer architecture Tolstikhin et al. (2021) with gated residual updates Jang et al. (2017). Both the inverse mode and the VQ-VAE bottlenecks are updated using an Adam optimizer Diederik et al. (2014) with a default learning rate of 0.0001 without weight decay.

A.2 Robotic Arm under Exogenous Observations

Using a robotic arm (Annin) with 6 degrees of freedom, there are 5 possible abstract actions: forward, reverse, left, right, and stay. The robot arm moves within 9 possible positions in a virtual 3x3 grid, with walls between some cells. The center of each cell is equidistant from adjoining cells. The end effector is kept at a constant height. We compute each cell's centroid and compose a transformation from the joint space of the robot to particular grid cells via standard inverse kinematics calculations. Two cameras are used to take still images. One camera is facing the front of the robot, and the other camera is facing down from above the robot. When a command is received, the robot moves from one cell center to another cell center, assuming no wall is present. After each movement, still images (640x480) are taken from two cameras and appended together into one image (1280x480). During training, only the forward facing, down-sampled (256x256) image is used. Each movement takes one second. After every 500 joint space movements, we re-calibrate the robot to the grid to avoid position drift.

We collected 6 hours (14000 data samples) of the robot arm following a uniformly random policy. There were no episodes or state resets. In addition to the robot, there are several distracting elements in the image. A looped video (`https://www.youtube.com/watch?v=zRpazyH1WzI`) plays on a large display in high resolution (4K video) at 2x speed. Four drinking toy birds, a color-changing lamp, and flashing streamer lights are also present. During the last half hour of image collection, the distracting elements are moved and/or removed to simulate additional uncertainty in the environment. An illustration of the setup is in Figure 7, along with the specific counter-example for one-step inverse models.

Latent State Visualizations: We learned a visualization of the latent state by learning a small convolutional neural network to map from the latent state $f(x_t)$ to an estimate $\hat{x}_t$ the observation x_t by optimizing the mean-square error reconstruction loss $||\hat{x} - x_t||^2$.

Videos of the latent state visualization for the baseline autoencoder and *AC-State* are included in the supplementary material. In each video, the frontal view (ground truth) is shown on the left, the top-down view (ground truth) is shown in the middle, the reconstruction of the frontal view from the latent state is shown on the right.

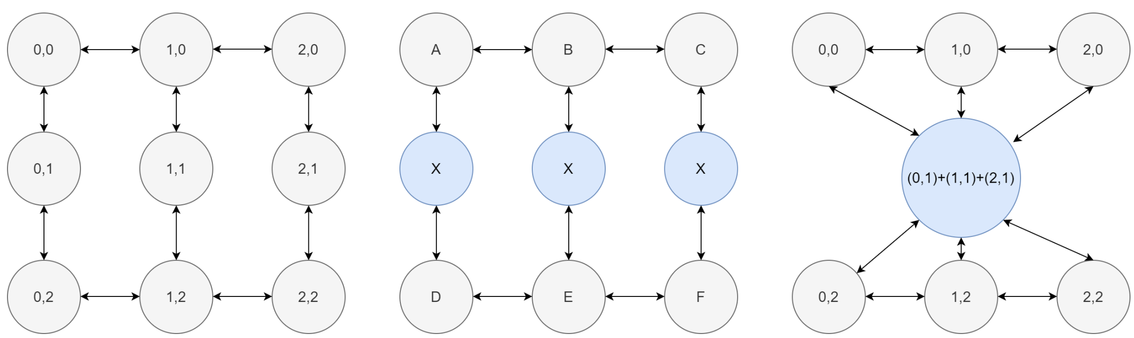

Figure 7: The robot arm has five actions and moves within nine possible control-endogenous states (left). The transition directions are indicated by the arrows. For example, if the robot arm is at (0,0) and selects the down action, it moves to (0,1), but if it selects the up action, it remains at the same position. A simple inverse model can achieve perfect accuracy even if the middle row of true control-endogenous states are mapped to a single latent state (middle). This leads the one-step inverse model to merge them (right).

A.3 Matterport Simulator with Exogenous Observations

We evaluated *AC-State* on the matterport simulator introduced in Chang et al. (2017). The simulator contains indoor houses in which an agent can navigate. The house contains a finite number of viewpoints which the agent can navigate to. At each viewpoint, the agent has control of its viewing angle (by turning left or right by an angle) and its elevation: in total there are 12 possible viewing angles per viewpoint and 3 possible elevations. We collect data using a random rollout policy. At each step of the rollout policy, the agent navigates to a neighbouring viewpoint. We also randomly change the agent elevation at some of steps of the rollout policy, in order to introduce exogenous information which the agent cannot control. We collect a single long episode of 20,000 state-transitions. The control-endogenous latent state in this setup is the viewpoint information while the exogenous information is the information regarding agent elevation.

Experimental Details: The model input is the panorama of the current viewpoint i.e. 12 images for the 12 possible views of each viewpoint. The `AC-State` model f is parameterized using a vision transformer

(ViT) Dosovitskiy et al. (2021). Each view within the panorama is fed separately into the ViT as a sequence of patches along with a learnable token called the class (or CLS) following the procedure in Dosovitskiy et al. (2021). To obtain the viewpoint representation, we take the representation corresponding to the CLS token of each view and take the mean across all views. We discretize this representation using a VQ-VAE bottleneck van den Oord et al. (2017) to obtain the final representation. We use a 6-layer transformer with 256 dimensions in the embedding. We use a feedforward network (FFN) after every attention operation in the ViT similar to Vaswani et al. (2017). The FFN is a 2 layer MLP with a GELU activation Hendrycks & Gimpel (2016) which first projects the input to a higher dimension D and then projects it back to the original dimension. We set the FFN dimension D to 512. We use 4 heads in the ViT. We train the model for 20 epochs using Adam optimizer Diederik et al. (2014) with learning rate 1e-4. The model is trained to predict the viewpoint of the next state as the action.

Results: We present the results for this experiment in Figure 6 (right). The *Controllable Latent State Accuracy* is the viewpoint prediction accuracy for the current state. The *Exogenous Noise-Ignoring Accuracy.* is calculate as $1 - \frac{\mathcal{E}-33.33}{100-33.33}$, where $\mathcal{E}$ is elevation prediction accuracy. Thus a higher elevation prediction accuracy leads to a lower the exogenous noise-inducing accuracy. We can see that the proposed *AC-State* model has the highest control-endogenous latent state and exogenous noise-ignoring accuracy. Thus, it outperforms the baselines we considered at capturing Control-endogenous Latent State information while ignoring exogenous noise. We calculated state parsimony as $\frac{\text{Num. Ground Truth States}}{\text{Num. Discovered States}}$. Therefore, a lower state parsimony denotes a high number of discovered states which means that the model fails at ignoring exogenous information. The proposed model has the highest state parsimony which shows the effectiveness of the model in ignoring the exogenous noise whilst only capturing control-endogenous latent state.

B Additional Discussion on Related Works

In this section, we provide a clear example of why an one-step inverse model might not be useful to ignore exogenous noise. The idea of using a simple one step inverse dynamics models have been explored in the past (Pathak et al., 2017; Bharadhwaj et al., 2022), yet the one step inverse model has counterexamples establishing that it fails to capture the full control-endogenous latent state (Efroni et al., 2022c; Misra et al., 2020; Rakelly et al., 2021; Hutter & Hansen, 2022). Intuitively, the 1-step inverse model is under-constrained and thus may incorrectly merge distinct states which are far apart in the MDP but have a similar local structure. As a simple example, suppose we have a cycle of states: $s_1, s_2, s_3, s_4, s_5, s_6$ where $a = 0$ moves earlier in the cycle and $a = 1$ moves later in the cycle. Suppose s_1, s_4 are merged into a distinct latent state s_i, s_2, s_5 are merged into a distinct latent state s_j and s_3, s_6 are merged into a distinct latent state s_k. The inverse-model examples are: $(s_i, s_j, 1), (s_j, s_i, 0), (s_j, s_k, 1), (s_k, s_j, 0)$, $(s_k, s_i, 1), (s_i, s_k, 0)$. Because all of these examples have distinct inputs, a 1-step inverse model still has zero error despite the incorrect merger of $\{s_1, s_4\}$, $\{s_2, s_5\}$, and $\{s_3, s_6\}$.

C Algorithm Details

We provide more detailed algorithm descriptions for `AC-State` with a random rollout policy (Algorithm 1) and with a planning policy (Algorithm 2). The latter uses a findgoal function which selects a low-count state using breadth-first search along with a plan function that finds an optimal action using Dijkstra's algorithm.

Algorithm 1 AC-State with Random Policy

Initialize observation trajectory x and action trajectory a. Initialize encoder f_θ. Assume a control-endogenous diameter of K and a number of samples to collect T, and a set of actions $\mathcal{A}$, and a number of training iterations N.
$x_1 \sim U(\mu(x))$
for t = 1, 2, ..., T **do**
  $a_t \sim U(\mathcal{A})$
  $x_{t+1} \sim \mathbb{P}(x'|x_t, a_t)$
for n = 1, 2, ..., N **do**
  $t \sim U(1, T)$ and $k \sim U(1, K)$
  $\mathcal{L} = \mathcal{L}_{\texttt{AC-State}}(f_\theta, t, x, a, k) + \mathcal{L}_{\text{Bottleneck}}(f_\theta, x_t) + \mathcal{L}_{\text{Bottleneck}}(f_\theta, x_{t+k})$
  Update θ to minimize $\mathcal{L}$ by gradient descent.

Algorithm 2 AC-State with Planning Policy

Initialize a replay buffer D. Initialize encoder f_θ. Assume a number of samples to collect T, a set of actions $\mathcal{A}$.
$x_1 \sim U(\mu(x))$, $a_1 \sim U(\mathcal{A})$, $t_g := 1$, and $\mathcal{T} = \{\}$
for t = 1, 2, ..., T **do**
  $x_{t+1} \sim \mathbb{P}(x'|x_t, a_t)$
  $s_t = f_\theta(x_t)$
  $s_{t+1} = f_\theta(x_{t+1})$
  Update tabular-MDP $\mathcal{M}$ with (s_t, a_t, s_{t+1}).
  if $t = t_g$ **then**
    Pick a new goal
    $t_s := t$
    $t_g, g := findgoal(s_t, \mathcal{M})$
    $a_t \sim U(\mathcal{A})$
    Add t to $\mathcal{T}$
  else
    $a_t := plan(s_t, g, \mathcal{M})$
  $K_t := t_g - t$
  $j \sim \mathcal{T}$ and $k \sim U(j, K_j)$
  $\mathcal{L} = \mathcal{L}_{\texttt{AC-State}}(f_\theta, j, x, a, k) + \mathcal{L}_{\text{Bottleneck}}(f_\theta, x_j) + \mathcal{L}_{\text{Bottleneck}}(f_\theta, x_{j+k})$
  Update θ to minimize $\mathcal{L}$ by gradient descent.

D Detailed Theory and Discussion

D.1 High-Level Overview of Theory

We present an asymptotic analysis of `AC-State` showing it recovers the control-endogenous latent state encoder $f_\star$. The mathematical model we consider is the deterministic Ex-BMDP. There, the transition model of the latent state decomposes into a control-endogenous latent state, which evolves deterministically, along with a noise term–the uncontrol-endogenous portion of the state. The noise term may be an arbitrary temporally correlated stochastic process. If the reward does not depend on this noise, any optimal policy may be expressed in terms of this control-endogenous latent state. In this sense, the recovered control-endogenous latent state is sufficient for achieving optimal behavior.

Intuitively, the Ex-BMDP is similar to a video game, in which a "game engine" takes player actions and keeps track of an internal game state (the control-endogenous state component), while the visuals and sound are rendered as a function of this compact game state. A modern video game's core state is often orders of magnitude smaller than the overall game.

The algorithm we propose for recovering the optimal control-endogenous latent state involves (i) an action prediction term; and (ii) a mutual information minimization term. The action prediction term forces the learned representation $\widehat{f}(x)$ to capture information about the dynamics of the system. At the same time, this representation for $\widehat{f}(x)$ (which is optimal for action-prediction) may also capture information which is unnecessary for control. In our analysis we assume that $\widehat{f}(x)$ has discrete values and show the control-endogenous latent state is the unique coarsest solution.

To enable more widespread adoption in deep learning applications, we can generalize this notion of coarseness to minimizing mutual information between x and $f(x)$. These are related by the data-processing inequality; coarser representation reduces mutual information with the input. Similarly, the notion of mutual information is general as it does not require discrete representation.

D.2 The Control-Endogenous Partition is a Bayes' Optimal Solution

Consider the generative process in which x is sampled from a distribution μ, the agent executes a policy π for t time steps and samples x'. Denote by $\mathbb{P}_{\pi,\mu}(x, x', t)$ as the joint probability, and by $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ as the probability that under this generative process the action upon observing x is a. The following result, which builds on Proposition 3, shows that the optimal bayes solution $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ is equal to $\mathbb{P}_{\pi,\mu}(a \mid f_\star(x), f_\star(x'), t)$ for $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$, where $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ is the probability to sample x .

Proposition 9. *Assume that π is and endogenous policy. Let $x \sim \mu$ for some distribution μ. Then, the Bayes' optimal predictor of the action-prediction model is piece-wise constant with respect to the control-endogenous partition: for all $a \in \mathcal{A}$, $t > 0$ and $x, x' \in \mathcal{X}$ such that $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ it holds that*

$$\mathbb{P}_{\pi,\mu}(a \mid x, x', t) = \mathbb{P}_{\pi,\mu}(a \mid f_\star(x), f_\star(x'), t).$$

We comment that the condition $\mathbb{P}_{\pi,\mu}(x, x', t) > 0$ is necessary since ,otherwise, the conditional probability $\mathbb{P}_{\pi,\mu}(a \mid x, x', t)$ is well not defined.

Proposition 9 is readily proved via the factorization of the future observation distribution to control-endogenous and exogenous parts that holds when the executed policy does not depend on the exogenous state (Proposition 3).

Proof. The proof follows by applying Bayes' theorem, Proposition equation 3, and eliminating terms from the numerator and denominator.

Fix any $t > 0$, $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$ such that $\mathbb{P}_{\pi,\mu}(x', x, t) > 0$. Let $s = f_\star(x)$, $s' = f_\star(x')$, $e = f_{\star,e}(x)$, and $e' = f_{\star,e}(x')$. The following relations hold.

$$\begin{aligned}
&\mathbb{P}_{\pi,\mu}(a \mid x', x, t) \\
&\stackrel{(a)}{=} \frac{\mathbb{P}_{\pi,\mu}(x' \mid x, a, t)\mathbb{P}_{\pi,\mu}(a \mid x)}{\sum_{a'} \mathbb{P}_{\pi,\mu}(x' \mid x, a', t)} \\
&\stackrel{(b)}{=} \frac{\mathbb{P}_{\pi,\mu}(x' \mid x, a, t)\pi(a \mid s)}{\sum_{a'} \mathbb{P}_{\pi,\mu}(x' \mid x, a', t)\pi(a' \mid s)} \\
&\stackrel{(c)}{=} \frac{q(x' \mid s', e')\mathbb{P}_{\pi,\mu}(s' \mid s, a, t)\mathbb{P}_{\pi,\mu}(e' \mid e, t)\pi(a \mid s)}{\sum_{a'} q(x' \mid s', e')\mathbb{P}_{\pi,\mu}(s' \mid s, a', t)\mathbb{P}_{\pi,\mu}(e' \mid e, t)\pi(a' \mid s)} \\
&= \frac{\mathbb{P}_{\pi,\mu}(s' \mid s, a, t)\pi(a \mid s)}{\sum_{a'} \mathbb{P}_{\pi,\mu}(s' \mid s, a', t)\pi(a' \mid s)}.
\end{aligned}$$

Relation (a) holds by Bayes' theorem. Relation (b) holds by the assumption that π is endogenous. Relation (c) holds by Proposition equation 3.

Thus, $\mathbb{P}_{\pi,\mu}(a \mid x', x, t) = \mathbb{P}_{\pi,\mu}(a \mid f_\star(x'), f_\star(x), t)$, and is constant upon changing the observation while fixing the control-endogenous latent state. □

D.3 The Coarsest Partition is the Control-Endogenous State Partition

Proposition 9 from previous section shows that the multi-step action-prediction model is piece-wise constant with respect to the partition induced by the control-endogenous states $f_\star : \mathcal{X} \to [S]$. In this section, we assume that the executed policy is an endogenous policy that induces sufficient exploration. With this, we prove that there is no coarser partition of the observation space such that the set of inverse models are piece-wise constant with respect to it.

We make the following assumptions on the policy by which the data is collected.

Assumption 10. *Let $T_{\mathcal{D}}(s' \mid s)$ be the Markov chain induced on the control-endogenous state space by executing the policy $\pi_{\mathcal{D}}$ by which AC-State collects the data.*

1. *The Markov chain $T_{\mathcal{D}}$ has a stationary distribution $\mu_{\mathcal{D}}$ such that $\mu_{\mathcal{D}}(s,a) > 0$ and $\pi_{\mathcal{D}}(a \mid s) \geq \pi_{\min}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*
2. *The policy $\pi_{\mathcal{D}}$ by which the data is collected reaches all accessible states from any states. For any $s, s' \in \mathcal{S}$ and any $h > 0$ if s' is reachable from s then $\mathbb{P}_{\mathcal{D}}(s' \mid s, h) > 0$.*
3. *The policy $\pi_{\mathcal{D}}$ does not depend on the exogenous state, and is an endogenous policy.*

See Levin & Peres (2017), Chapter 1, for further discussion on the classes of Markov chains for which the assumption on the stationary distribution hold. The second and third assumptions are satisfied for the the random policy that simply executes random actions.

We consider the stochastic process in which an observation is sampled from a distribution μ such that $\mu(s) = \mu_{\mathcal{D}}(s)$ for all s. Then, the agent executes the policy $\pi_{\mathcal{D}}$ for t time steps. For brevity, we denote the probability measure induced by this process as $\mathbb{P}_{\mathcal{D}}$.

We begin by defining several useful notions. We denote the set of reachable control-endogenous states from s in h time steps as $\mathcal{R}_h(s)$.

Definition 11 (Reachable Control-Endogenous States)**.** *Let the set of reachable control-endogenous states from $s \in \mathcal{S}$ in $h > 0$ time steps be $\mathcal{R}(s,h) = \{s' \mid \max_\pi \mathbb{P}_\pi(s' \mid s, h) = 1\}$.*

Observe that every reachable state from s in h time steps satisfies that $\max_\pi \mathbb{P}_{\pi,\mu}(s' \mid s_0 = s, h) = 1$ due to the deterministic assumption of the control-endogenous dynamics.

Next, we define a notion of *consistent partition* with respect to a set of function values. Intuitively, a partition of space $\mathcal{X}$ is consistent with a set of function values if the function is piece-wise constant on that partition.

Definition 12 (Consistent Partition with respect to $\mathcal{G}$)**.** *Consider a set $\mathcal{G} = \{g(a,y,y')\}_{y,y' \in \mathcal{Y}, a \in \mathcal{A}}$ where $g : \mathcal{A} \times \mathcal{Y} \times \mathcal{Y} \to [0,1]$. We say that $f : \mathcal{Y} \to [N]$ is a* consistent partition *with respect to $\mathcal{G}$ if for all $y, y_1', y_2' \in \mathcal{Y}$, $f(y_1') = f(y_2')$ implies that $g(a,y,y_1') = g(a,y,y_2')$ for all $a \in \mathcal{A}$.*

Observe that Proposition 9 shows that the partition of $\mathcal{X}$ according to $f_\star$ is consistent with respect to $\{\mathbb{P}_{\mathcal{D}}(a \mid x, x', h) \mid x, x' \in \mathcal{X}, h \in [H] \text{ s.t. } \mathbb{P}_{\mathcal{D}}(x, x', h) > 0\}$, since, by Proposition 9, $\mathbb{P}_{\mathcal{D}}(a \mid x, x', h) = \mathbb{P}_{\mathcal{D}}(a \mid f_\star(x), f_\star(x'), h)$.

Towards establishing that the coarsest abstraction according to the `AC-State` objective is $f_\star$ we make the following definition.

Definition 13 (The Generalized Inverse Dynamics Set $\mathrm{AC}(s,h)$)**.** *Let $s \in \mathcal{S}, h \in \mathbb{N}$. We denote by $\mathrm{AC}(s,h)$ as the set of multi-step inverse models accessible from s in h time steps. Formally,*

$$\mathrm{AC}(s,h) = \{\mathbb{P}_{\mathcal{D}}(a \mid s', s'', h') : s' \in \mathcal{R}(s, h - h'), s'' \in \mathcal{R}(s', h'), a \in \mathcal{A}, h' \in [h]\}. \tag{5}$$

Observe that in equation equation 5 the inverse function $\mathbb{P}_{\mathcal{D}}(a \mid s', s'', h')$ is always well defined since $\mathbb{P}_{\mathcal{D}}(s', s'', h') > 0$. It holds that

$$\mathbb{P}_{\mathcal{D}}(s', s'', h') = \mathbb{P}_{\mathcal{D}}(s')\mathbb{P}_{\mathcal{D}}(s'' \mid s', h') > 0,$$

since $\mathbb{P}_{\mathcal{D}}(s') > 0$ and $\mathbb{P}_{\mathcal{D}}(s'' \mid s', h') > 0$. The inequality $\mathbb{P}_{\mathcal{D}}(s') > 0$ holds by the assumption that the stationary distribution when following $\mathcal{U}$ has positive support on all control-endogenous states (Assumption 10). The inequality $\mathbb{P}_{\mathcal{D}}(s'' \mid s', h') > 0$ holds since, by definition $s'' \in \mathcal{R}(s, h')$ is reachable from s' in h' time steps; hence, $\mathbb{P}_{\mathcal{D}}(s'' \mid s', h') > 0$ by the fact that Assumption 10 implies that the policy $\pi_{\mathcal{D}}$ induces sufficient exploration.

Theorem 14 ($f_\star$ is the coarsest partition consistent with *AC-State* objective)**.** *Assume 2 and 10 holds. Then there is no coarser partition than $f_\star$ which is consistent with* $\mathrm{AC}(s, D)$ *for any $s \in \mathcal{S}$.*

Proof. We will show inductively that for any $h > 0$ and $s \in \mathcal{S}$ there is no coarser partition than $\mathcal{R}(s, h)$ for the set $\mathcal{R}(s, h)$ that is consistent with $\mathrm{AC}(s, h)$. Since the set of reachable states in $h = D$ time steps is $\mathcal{S}$–all states are reachable from any state in D time steps–it will directly imply that there is no coarser partition than $\mathcal{R}(s, D) = \mathcal{S}$ consistent $\mathrm{AC}(s, D)$.

Base case, $h = 1$. Assume that $h = 1$ and fix some $s \in \mathcal{S}$. Since the control-endogenous dynamics is deterministic, there are A reachable states from s. Observe the inverse dynamics for any $s' \in \mathcal{R}(s, 1)$ satisfies that

$$\mathbb{P}_{\mathcal{D}}(a \mid s, s', 1) = \begin{cases} 1 & \text{if } a \text{ leads } s' \text{ from } s \\ 0 & \text{o.w.} \end{cases}. \tag{6}$$

This can be proved by an application of Bayes' rule:

$$\begin{aligned}\mathbb{P}_{\mathcal{D}}(a \mid s, s', 1) &= \frac{\mathbb{P}_{\mathcal{D}}(s' \mid s = s, a, 1)\pi_{\mathcal{D}}(a \mid s)}{\sum_{a'} \mathbb{P}_{\mathcal{D}}(s' \mid s = s, a', 1)\pi_{\mathcal{D}}(a' \mid s)} \\ &= \frac{T(s' \mid s, a)\pi_{\mathcal{D}}(a \mid s)}{\sum_{a'} T(s' \mid s, a')\pi_{\mathcal{D}}(a' \mid s)} \\ &\begin{cases} \geq \pi_{\min} & (s, a) \text{ leads to } s' \\ = 0 & o.w. \end{cases},\end{aligned}$$

where the last relation holds by Assumption 10. Furthermore, observe that since $s' \in \mathcal{R}(s, 1)$, i.e., it is reachable from s, the probability function $\mathbb{P}_{\mathcal{D}}(a \mid s, s', 1)$ is well defined.

Hence, by equation equation 6, we get that for any $s_1', s_2' \in \mathcal{R}(s, 1)$ such that $s_1' \neq s_2'$ it holds that exists $a \in \mathcal{A}$ such that

$$\pi_{\min} \geq \mathbb{P}_{\mathcal{D}}(a \mid s, s_1', 1) \neq \mathbb{P}_{\mathcal{D}}(a \mid s, s_2', 1) = 0.$$

Specifically, choose a such that taking a from s leads to s_1' and see that, by equation equation 6,

$$\pi_{\min} \geq \mathbb{P}_{\mathcal{D}}(a \mid s, s_1', 1) \neq \mathbb{P}_{\mathcal{D}}(a \mid s, s_2', 1) = 0.$$

Lastly, by the fact that $s \in \mathcal{S}$ is an arbitrary state, the induction base case is proved for all $s \in \mathcal{S}$.

Induction step. Assume the induction claim holds for all $t \in [h]$ where $h \in \mathbb{N}$. We now prove it holds for $t = h + 1$.

Fix some $s \in \mathcal{S}$. We prove the induction step and show that $\mathcal{R}(s, h + 1)$ is the coarsest partition which is consistent $\mathrm{AC}(s, h + 1)$. Meaning, there exists $\tilde{s}, t \in [h + 1], a$ such that $\mathbb{P}(a \mid \tilde{s}, s_1', 1), \mathbb{P}(a \mid \tilde{s}, s_1', 1) \in \mathrm{AC}(\bar{s}, h + 1)$ and

$$\mathbb{P}_{\mathcal{D}}(a \mid \tilde{s}, s_1', t) \neq \mathbb{P}_{\mathcal{D}}(a \mid \tilde{s}, s_2', t). \tag{7}$$

Observe that, by Definition 13, it holds that,

$$\mathrm{AC}(s, h + 1) = \{\mathbb{P}_{\mathcal{D}}(a \mid s, s', h + 1)\}_{s' \in \mathcal{R}(s, h+1)} \cup_{\bar{s} \in \mathcal{R}(s,1)} \mathrm{AC}(\bar{s}, h).$$

Meaning, the set $\mathrm{AC}(s, h+1)$ can be written as the union of (1) the set $\{\mathbb{P}_{\mathcal{D}}(a \mid s, s', h+1)\}_{s' \in \mathcal{R}(s,h+1)}$, and (2) the union of the sets $\mathrm{AC}(\bar{s}, h)$ for all $\bar{s}$ which is reachable from s in a single time step.

By the induction hypothesis, the coarsest partition which is consistent with $\mathrm{AC}(\bar{s}, h)$ is $\cup_{h'=1}^{h} \mathcal{R}(\bar{s}, h')$. We only need to prove, that for any $\bar{s}_1, \bar{s}_2 \in \mathcal{R}(s, h+1)$ such that $\bar{s}_1 \neq \bar{s}_2$ exists some $a \in \mathcal{A}, h' \in [h]$ and $s_{h'} \in \mathcal{R}(s, h+1)$ such that

$$\mathbb{P}_{\mathcal{D}}(a \mid s_{h'}, \bar{s}_1, h') \neq \mathbb{P}_{\mathcal{D}}(a \mid s_{h'}, \bar{s}_2, h'),$$

this will imply that the set of reachable states in $h+1$ time states is also the coarsest partition which is consistent with $\mathrm{AC}(s, h+1)$.

Fix $\bar{s}_1, \bar{s}_2 \in \mathcal{R}(s, h+1)$ such that $\bar{s}_1 \neq \bar{s}_2$ we show that exists a certificate in $\mathrm{AC}(s, h+1)$ that differentiate between the two by considering three cases.

1. **Case 1:** Both $\bar{s}_1$ and $\bar{s}_2$ are reachable from all $s' \in \mathcal{R}(s, 1)$. In this case, for all $s' \in \mathcal{R}(s, 1)$ it holds that $\bar{s}_1, \bar{s}_2 \in \mathcal{R}(s', h)$. By the induction hypothesis, $\bar{s}_1$ and $\bar{s}_2$ cannot be merged while being consistent with $\mathrm{AC}(s', h)$.

2. **Case 2:** Exists $s' \in \mathcal{R}(s, 1)$ such $\bar{s}_1$ is reachable from s' in h time steps and $\bar{s}_2$ is not. Let a be the action that leads to s' from state s. In that case, it holds by the third assumption of Assumption 10 that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}(a \mid s, \bar{s}_1, h+1) &\overset{(a)}{=} \frac{\mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s, a, h+1)\pi_{\mathcal{D}}(a \mid s)}{\sum_{a'} \mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s, a', h+1)\pi_{\mathcal{D}}(a' \mid s)} \\
&\overset{(b)}{=} \frac{\mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s', h)\pi_{\mathcal{D}}(a \mid s)}{\sum_{a'} \mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s, a', h+1)\pi_{\mathcal{D}}(a' \mid s)} \\
&\overset{(c)}{\geq} \pi_{\min} \frac{\mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s', h)}{\sum_{a'} \mathbb{P}_{\mathcal{D}}(\bar{s}_1 \mid s, a', h+1)} \\
&\overset{(d)}{>} 0.
\end{aligned} \tag{8}$$

   Relation (a) holds by Bayes' rule. Relation (b) holds by the fact that (s, a) determinstically leads to s'. Relation (c) and (d) holds by Assumption 10.

   Observe that $\mathbb{P}_{\mathcal{D}}(a \mid s, \bar{s}_2, h+1) = 0$ since $\bar{s}_2$ is not reachable upon taking action a from state s, by the assumption. Combining this fact with equation equation 8 implies that

$$0 < \mathbb{P}_{\mathcal{D}}(a \mid s, \bar{s}_1, h+1) \neq \mathbb{P}_{\mathcal{D}}(a \mid s, \bar{s}_2, h+1) = 0.$$

   Hence, exists a certificate that differentiates between $\bar{s}_1$ and $\bar{s}_2$. Observe that since $\bar{s}_2 \in \mathcal{R}(s, h+1)$, i.e., it is reachable from s, it holds that $\mathbb{P}_{\mathcal{D}}(a \mid s, \bar{s}_2, h+1) = 0$, i.e., it is well defined.

3. **Case 3:** Exists $s' \in \mathcal{R}(s, 1)$ such $\bar{s}_2$ is reachable from s' in h time steps and $\bar{s}_1$ is not. Symmetric to case 2.

This establishes the result we needed to show in equation equation 7 and, hence, the induction and result hold. □

E Discussion: What does Control-Endogenous Latent State Capture? Does it ignore Task-Relevant and Reward-Relevant Information?

The control-endogenous latent state is defined by having factorized dynamics that depend on actions, while the exogenous dynamics do not depend on actions. If we refer to the control-endogenous state as s and the exogenous state as e, we can write the factorized dynamics as: $\mathbb{P}(s'|s, a)\mathbb{P}(e'|e)$. In particular, we want to find the smallest s such that the latent dynamics follow this factorization.

Causal Dynamics Learning (Wang et al., 2022b) provided some definitions that are helpful for building intuitions about the semantics of the control-endogenous state. In their work, they decompose a factorized state into controllable factors, action-relevant factors, and action-irrelevant factors. They define controllable factors as any causal descendants of actions or other controllable factors. They define action-relevant factors as causal parents of controllable factors or other action-relevant factors. All other factors are defined as action-irrelevant. While `AC-State` learns a general encoder instead of relying on a given factorization, the special case of known factors can be used to build intuition.

An example is given in Wang et al. (2022b) involving six known factors: $z_1, z_2, z_3, z_4, z_5, z_6$. The dynamics are factorized as: $\mathbb{P}(z_1'|z_1, a)\mathbb{P}(z_2'|z_1, z_2, z_3)\mathbb{P}(z_3'|z_3, z_4)\mathbb{P}(z_4'|z_4)\mathbb{P}(z_5'|z_5, z_6)\mathbb{P}(z_6'|z_6)$. We can assign these factors to either the control-endogenous state or exogenous state. The smallest way to define the control-endogenous state is: $s = (z_1, z_2, z_3, z_4)$, while the exogenous state is set to $e = (z_5, z_6)$. Intriguingly, the control-endogenous state includes many factors which are not controllable.

As a simple example, we can imagine a robotic arm that is allowed to interact with two blocks (one red and one blue). What is the control-endogenous latent state for this environment? It clearly includes the robotic arm itself, as it can be manipulated using the actions. Likewise, the physical properties of the blocks (their position and weight) will be included in the control-endogenous latent state.

The color of the blocks is not part of the control-endogenous latent state. What if our task is to retrieve the red block? In general, the control-endogenous latent state will not be sufficient for solving the task. If task completion has a causal effect on the rest of the control-endogenous state, then the task relevant information is control-endogenous. As a concrete example, if the agent were forced to redo the task of retrieving the red block until successful, then the color of the blocks would become control-endogenous. Intuitively, in the most natural and realistic environments, rewards and task completion will have a causal effect on the rest of the control-endogenous state. On the other hand, if the episodes are terminated immediately after a reward is received, it would be useful to add reward prediction as an auxiliary task for learning the representation.