A. Notation

We provide a summary of key notation used throughout the paper here.

$\mathbf{PA}_{\mathcal{G}}(X)$: the parents of node X in the causal graph $\mathcal{G}$. When $\mathcal{G}$ is clear from the setting, abbreviate this notation to $\mathbf{PA}(X)$.
$\mathbf{AN}_{\mathcal{G}}(X)$: the ancestors of node X in $\mathcal{G}$ (again, $\mathcal{G}$ omitted when unambiguous).
$[x]_S : [x_{i_1}, \ldots, x_{i_k} | i_j \in S]$
π_M : the stationary distribution given by some fixed policy in an MDP M.
q : the emission function of a block MDP.
$\mathcal{E}$: a set of environments.

B. Proofs

Technical notes and assumptions. In order for the block MDP assumption to be satisfied, we will require that the interventions defining each environment only occur outside of the causal ancestors of the reward. Otherwise, the different environments will have different latent state dynamics, which violates our assumption that the environments are obtained by a noisy emission function from the latent state space $\mathcal{S}$. Although ICP will still find the correct causal variables in this setting, this state abstraction will no longer be a model irrelevance state abstraction over the union of the environments.

Theorem 1. *Consider a family of MDPs $M_{\mathcal{E}} = \{(\mathcal{X}, A, R, P_e, \gamma) | e \in \mathcal{E}\}$, with $\mathcal{X} = \mathbb{R}^k$. Let $M_{\mathcal{E}}$ satisfy Assumptions 1-3. Let $S_R \subseteq \{1, \ldots, k\}$ be the set of variables such that the reward $R(x, a)$ is a function only of $[x]_{S_R}$ (x restricted to the indices in S_R). Then let $S = \mathbf{AN}(R)$ denote the ancestors of S_R in the (fully observable) causal graph corresponding to the transition dynamics of $M_{\mathcal{E}}$. Then the state abstraction $\phi_S(x) = [x]_S$ is a model-irrelevance abstraction for every $e \in \mathcal{E}$.*

Proof. To prove that ϕ_S is a model-irrelevance abstraction, we must first show that $r(x) = r(x')$ for any $x, x' : \phi_S(x) = \phi_S(x')$. For this, we note that $\mathbb{E}[R(x)] = \int_{r \in \mathbb{R}} r dp(r|x) = \int_{r \in \mathbb{R}} r dp(r|[x]_S, [x]_{S^C})$ and, because by definition $S^C \subset \mathbf{PA}(R)^C$, we have that $R \perp [x]_{S^C}$. Therefore,

$$\mathbb{E}[R(x)] = \int_{r \in \mathbb{R}} r dp(r|[x]_S) = \int_{r \in \mathbb{R}} r dp(r|[x']_S) = \mathbb{E}[R(x')]. \tag{7}$$

To show that $[x]_S$ is a MISA, we must also show that for any x_1, x_2 such that $\phi(x_1) = \phi(x_2)$, and for any $e \in \mathcal{E}$, the distribution over next state equivalence classes will be equal for x_1 and x_2.

$$\sum_{x' \in \phi^{-1}(\bar{X})} P^e_{x_1 x'} = \sum_{x' \in \phi^{-1}(\bar{X})} P^e_{x_2 x'}.$$

For this, it suffices to observe that S is closed under taking parents in the causal graph, and that by construction environments only contain interventions on variables outside of the causal set. Specifically, we observe that the probability of seeing any particular equivalence class $[x']_S$ after state x is only a function of $[x]_S$.

$$P([x']_S | x) = f([x]_S, [x']_S)$$

This allows us to define a natural decomposition of the transition function as follows.

$$P(x'|x) = P\left([x]_S \oplus [x]_{S^C} \middle| [x']_S \oplus [x']_{S^C}\right) \text{ which by the independent noise assumption gives}$$
$$P(x'|x) = f([x']_S, [x]_S) P([x']_{S^c} | x)$$

We further observe that since the components of x are independent, $\sum_{[x']_{S^C}} P([x']_{S^C}|x) = 1$. We now return to the

property we want to show:

$$\begin{aligned}\sum_{x' \in \phi^{-1}(\bar{x})} P^e_{x_1 x'} &= \sum_{x' \in \phi^{-1}(\bar{x})} f([x_1]_S, [x']_S) P(x'|x_1) \\ &= f(\phi(x_1), \bar{x}) \sum_{[x']_{S^C}} P\left([x']_{S^C} \middle| x_1\right) \\ &= f(\phi(x_1), \bar{x})\end{aligned}$$

and because $\phi(x_1) = \phi(x_2)$, we have

$$= f(\phi(x_2), \bar{x})$$

for which we can apply the previous chain of equalities backward to obtain

$$= \sum_{x' \in \phi^{-1}(\bar{x})} P^e_{x_2 x'}$$

□

Proposition 1 (Identifiability and Uniqueness of Causal State Abstraction)**.** *In the setting of the previous theorem, assume the transition dynamics and reward are linear functions of the current state. If the training environment set $\mathcal{E}_{train}$ satisfies any of the conditions of Theorem 2 (Peters et al., 2016) with respect to each variable in **AN(R)**, then the causal feature set ϕ_S is identifiable. Conversely, if the training environments don't contain sufficient interventions, then it may be that there exists a ϕ such that ϕ is a model irrelevance abstraction over $\mathcal{E}_{train}$, but not over $\mathcal{E}$ globally.*

Proof. The proof of the first statement follows immediately from the iterative application of the identifiability result of Peters et al. (2016) to each variable in the causal variables set.

For the converse, we consider a simple counterexample in which one variable x_m is constant in every training environment, with value v_m. Then letting $S = \mathbf{AN}(R)$, we observe that $S \cup \{m\}$ is also a model-irrelevance state abstraction.

First, we show $r(x_1) = r(x_2)$ for any $x_1, x_2 : \phi_{S\cup\{m\}}(x_1) = \phi_{S\cup\{m\}}(x_2)$.

$$\begin{aligned}p(R|x_1, a) &= p(R|x_1|_S, a) \\ &= p(R|x_1|_{S\cup\{m\}}, a, m = v_m) \\ &= p(R|(x_2|_{S\cup\{m\}}, a, m = v_m) \\ &= p(R|x_2, a)\end{aligned}$$

Finally, we must show that

$$\sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{X})} P_{x_1 x'} = \sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{X})} P_{x_2 x'}.$$

Again starting from the result of Theorem 1 we have:

$$\begin{aligned}\sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{x})} P_{x_1 x'} &= \sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{x})} f(x_1|_{S\cup\{m\}}, x'|_{S\cup\{m\}}) p(x'|x_1|_{(S\cup\{m\})^C}, m = v_m) \\ &= f(\phi_{S\cup\{m\}}(x_1), \bar{x}) \sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{x})} p(x'|x_1, m = v_m) \\ &= f(\phi_{S\cup\{m\}}(x_1), \bar{x})\end{aligned}$$

and because $\phi_{S\cup\{m\}}(x_1) = \phi_{S\cup\{m\}}(x_2)$, we have

$$= f(\phi_{S\cup\{m\}}(x_2), \bar{x})$$

for which we can apply the previous chain of equalities backward to obtain

$$= \sum_{x' \in \phi^{-1}_{S\cup\{m\}}(\bar{x})} P_{x_2 x'}$$

However, if one of the test environments contains the intervention $x_m \leftarrow v_m + \mathcal{N}(0, \sigma^2)$, then the distribution over next-states in the new environment will violate the conditions for a model-irrelevance abstraction. □

Theorem 2. *Consider an MDP M, with M' denoting a coarser bisimulation of M. Let ϕ denote the mapping from states of M to states of M'. Suppose that the dynamics of M are L-Lipschitz w.r.t. $\phi(X)$ and that T is some approximate transition model satisfying $\max_s \mathbb{E}\|T(\phi(s)) - \phi(T_M(s))\| < \delta$, for some $\delta > 0$. Let $W_1(\pi_1, \pi_2)$ denote the 1-Wasserstein distance. Then*

$$\mathbb{E}_{x\sim M'}[\|T(\phi(x)) - \phi(T_{M'}(x))\|] \leq \delta + 2LW_1(\pi_{\phi(M)}, \pi_{\phi(M')}). \tag{8}$$

We will use the shorthand π for $\pi_{\phi(M)}$, the distribution of state embeddings $\phi(M)$ corresponding to the behaviour policy, and π' for $\pi_{\phi(M')}$ for the distribution of state embeddings $\phi(M')$ given by the behaviour policy.

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim M'}[\|T(\phi(x)) - \phi(T_{M'}(x))\|] &= \mathbb{E}_{x\sim M'}[\min_{y\in X_M} \|T(\phi(x)) - T(\phi(y)) + T(\phi(y)) - \phi(T_M(x))\|] \\
&\leq \mathbb{E}_{x\sim M'}[\min_{y\in X_M} \|T(\phi(x)) - T(\phi(y))\| \\
&\quad + \|T(\phi(y)) - \phi(T_M(y))\| + \|\phi(T_M(y)) - \phi(T_M(x))\|]
\end{aligned}$$

Let γ be a coupling over the distributions of $\phi(M')$ and $\phi(M)$ such that $\mathbb{E}_{\gamma(\phi(x),\phi(y))}\|\phi(x) - \phi(y)\| = W_1(\pi, \pi')$

$$\begin{aligned}
&\leq \mathbb{E}_{x\sim M'}[\mathbb{E}_{\gamma(\phi(y)|\phi(x))}\|T(\phi(x)) - T(\phi(y))\|] + \delta + L\|x - y\|] \\
&\leq \mathbb{E}_{x\sim M'}[\mathbb{E}_{\gamma(\phi(y)|\phi(x))}L\|\phi(x) - \phi(y)\| + \delta + L\|\phi(x) - \phi(y)\|] \\
&= \mathbb{E}_{\gamma(\phi(x),\phi(y))}[L\|\phi(x) - \phi(y)\| + \delta + L\|\phi(x) - \phi(y)\|] \\
&= 2LW_1(\pi, \pi') + \delta
\end{aligned}$$

□

Theorem 4 (Existence of model-irrelevance state abstractions)**.** *Let $\mathcal{E}$ denote some family of bisimilar MDPs with joint state space $\mathcal{X}_\mathcal{E} = \cup_{e\in\mathcal{E}} X_e$. Let the mapping from states in M_e to the underlying abstract MDP $\bar{M}$ be denoted by f_e. Then if the states in $X_\mathcal{E}$ satisfy $x \in X_{e'} \cap X_e \implies f_e(x) = f_{e'}(x)$, then $\phi = \cup f_e$ is a model-irrelevance state abstraction for $\mathcal{E}$.*

Proof. First, note that $\cup f_e$ is well-defined (because each f agrees with the rest on the value of all states appearing in multiple tasks). Then ϕ will be a model-irrelevance abstraction for every MDP M_e because it agrees with f_e (a model-irrelevance abstraction). □

Theorem 3. *Let M be our block MDP and $\bar{M}$ the learned invariant MDP with a mapping $\phi : \mathcal{X} \mapsto \mathcal{Z}$. For any L-Lipschitz valued policy π the value difference is bounded by*

$$|Q^\pi(x, a) - \bar{Q}^\pi(\phi(x), a)| \leq \frac{J_R^\infty + \gamma L J_D^\infty}{1 - \gamma}. \tag{9}$$

Proof.

$$
\begin{aligned}
&\sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} |Q^\pi(x_t,a_t) - \bar{Q}^\pi(\phi(x_t),a_t)| \\
&\quad\le \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} |R(\phi(x_t),a,\phi(x_{t+1})) - r(x,a)| + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} |\mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)} V^\pi(x_{t+1}) - \mathbb{E}_{z_{t+1}\sim f(\cdot|\phi(x_t),a_t)} \bar{V}^\pi(z_{t+1})| \\
&\quad= J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|\mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)}[V^\pi(x_{t+1}) - \bar{V}^\pi(\phi(x_{t+1}))] + \mathbb{E}_{\substack{x_{t+1}\sim P(\cdot|x_t,a_t)\\ z_{t+1}\sim f(\cdot|\phi(x_t),a_t)}} [\bar{V}^\pi(\phi(x_{t+1})) - \bar{V}^\pi(z_{t+1})]\Big| \\
&\quad\le J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|\mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)}[V^\pi(x_{t+1}) - \bar{V}^\pi(\phi(x_{t+1}))]\Big| \\
&\qquad\quad + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|\mathbb{E}_{\substack{x_{t+1}\sim P(\cdot|x_t,a_t)\\ z_{t+1}\sim f(\cdot|\phi(x_t),a_t)}} [\bar{V}^\pi(\phi(x_{t+1})) - \bar{V}^\pi(z_{t+1})]\Big| \\
&\quad\le J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|\mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)}[V^\pi(x_{t+1}) - \bar{V}^\pi(\phi(x_{t+1}))]\Big| + \gamma L \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} W(\phi(P(\cdot|x_t,a_t)), f(\cdot|\phi(x_t),a_t)) \\
&\quad= J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|\mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)}[V^\pi(x_{t+1}) - \bar{V}^\pi(\phi(x_{t+1}))]\Big| + \gamma L J_D^\infty \\
&\quad\le J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \mathbb{E}_{x_{t+1}\sim P(\cdot|x_t,a_t)}\Big|[V^\pi(x_{t+1}) - \bar{V}^\pi(\phi(x_{t+1}))]\Big| + \gamma L J_D^\infty \\
&\quad\le J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|[V^\pi(x_t) - \bar{V}^\pi(\phi(x_t))]\Big| + \gamma L J_D^\infty \\
&\quad\le J_R^\infty + \gamma \sup_{x_t\in\mathcal{X},a_t\in\mathcal{A}} \Big|[Q^\pi(x_{t-1},a_{t-1}) - \bar{Q}^\pi(\phi(x_{t-1}),a_{t-1})]\Big| + \gamma L J_D^\infty \\
&\quad= \frac{J_R^\infty + \gamma L J_D^\infty}{1-\gamma}
\end{aligned}
$$

□

Proposition 2 (Lower bound on abstraction error)**.** *Let f_e be a mapping from $\mathcal{S}\to\mathcal{X}$. Fix some arbitrary policy ρ and let $v(s)$ denote the value of state s under ρ, with π its stationary distribution. If $\exists\, e, e', s, s'$ such that $f_e(s) = f_{e'}(s')$ (i.e. different states induce the same observation), then the following bound is a lower bound on the error obtained by a joint state abstraction over all environments.*

$$
\min_{\hat{v}} \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \textit{err}(\phi(X_e),\hat{v}) \ge \min_{s,s':v(s)\neq v(s')} \Big(|v(s)-v(s')|\Big) P_{\mathcal{E}}\Big((\phi(x)\neq f_e^{-1}(x)\Big) \ge \delta \frac{H(V(S)|X)-1}{\log|V(S)|} \tag{10}
$$

Where

$$
\textit{err}(\phi(X_e),\hat{v}) := \mathbb{E}_{\pi(X_e)}|\hat{v}(\phi(x)) - v(f_e^{-1}(x))|
$$

and

$$
\delta = \min_{s,s':v(s)\neq v(s')} \Big(|v(s)-v(s')|\Big)
$$

Proof. (Sketch) The error obtained by state abstraction will be at least the decoding error of values from abstract states scaled by δ. This in turn depends on how effectively it is possible to decode a potentially lossy mapping from observations back to states. This leads to the second inequality, due to Fano, where the entropy $H(V(S)|X)$ is given by marginalizatiion with respect to $v(s)$ of the following probability distributions.

$$
\begin{aligned}
p(x) &= \frac{1}{|\mathcal{E}|}\sum_{s,e} \mathbb{1}[f_e(s)=x]\pi(s) \\
p(s|x) &= \frac{1}{p(x)}\frac{1}{|\mathcal{E}|}\sum_{e}\pi(s)
\end{aligned}
$$

□

C. Implementation Details

C.1. Model Learning: Rich Observations

For the model learning experiments we use an almost identical encoder architecture as in Tassa et al. (2018), with two more convolutional layers to the convnet trunk. Secondly, we use `ReLU` activations after each convolutional layer, instead of `ELU`. We use kernels of size 3×3 with 32 channels for all the convolutional layers and set stride to 1 everywhere, except of the first convolutional layer, which has stride 2. We then take the output of the convolutional net and feed it into a single fully-connected layer normalized by `LayerNorm` (Ba et al., 2016). Finally, we add `tanh` nonlinearity to the 50 dimensional output of the fully-connected layer.

The decoder consists of one fully-connected layer that is then followed by four deconvolutional layers. We use `ReLU` activations after each layer, except the final deconvolutional layer that produces pixels representation. Each deconvolutional layer has kernels of size 3×3 with 32 channels and stride 1, except of the last layer, where stride is 2.

The dynamics and reward models are all MLPs with two hidden layers with 200 neurons each and `ReLU` activations.

C.2. Reinforcement Learning

For the reinforcement learning experiments we modify the Soft Actor-Critic PyTorch implementation by Yarats and Kostrikov (2020) and augment with a shared encoder between the actor and critic, the general model f_s and task-specific models f_η^e. The forward models are multi-layer perceptions with ReLU non-linearities and two hidden layers of 200 neurons each. The encoder is a linear layer that maps to a 50-dim hidden representation. We also use L1 regularization on the S latent representation. We add two additional dimensions to the state space, a spurious correlation dimension that is a multiplicative factor of the last dimension of the ground truth state, as well as an environment id. We add Gaussian noise $\mathcal{N}(0, 0.01)$ to the original state dimension, similar to how Arjovsky et al. (2019) incorporate noise in the label to make the task harder for the baseline.

Soft Actor Critic (SAC) (Haarnoja et al., 2018) is an off-policy actor-critic method that uses the maximum entropy framework to derive soft policy iteration. At each iteration, SAC performs soft policy evaluation and improvement steps. The policy evaluation step fits a parametric soft Q-function $Q(x_t, a_t)$ using transitions sampled from the replay buffer $\mathcal{D}$ by minimizing the soft Bellman residual,

$$J(Q) = \mathbb{E}_{(x_t, x_t, r_t, x_{t+1}) \sim \mathcal{D}} \left[\left(Q(x_t, a_t) - r_t - \gamma \bar{V}(x_{t+1}) \right)^2 \right].$$

The target value function $\bar{V}$ is approximated via a Monte-Carlo estimate of the following expectation,

$$\bar{V}(x_{t+1}) = \mathbb{E}_{a_{t+1} \sim \pi} \left[\bar{Q}(x_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1} | x_{t+1}) \right],$$

where $\bar{Q}$ is the target soft Q-function parameterized by a weight vector obtained from an exponentially moving average of the Q-function weights to stabilize training. The policy improvement step then attempts to project a parametric policy $\pi(a_t | x_t)$ by minimizing KL divergence between the policy and a Boltzmann distribution induced by the Q-function, producing the following objective,

$$J(\pi) = \mathbb{E}_{x_t \sim \mathcal{D}} \Big[\mathbb{E}_{a_t \sim \pi} [\alpha \log(\pi(a_t | x_t)) - Q(x_t, a_t)] \Big].$$

We provide the hyperparameters used for the RL experiments in Table 1.

C.3. IRM Hyperparameter Sweep

We found IRM to be very brittle, even on the original colored MNIST task they presented. We implement the same penalty and learning rate schedule proposed in their paper (Arjovsky et al., 2019), but found that we required a much smaller penalty weight to work. In Figure 8 we show the hyperparameter sweep we performed on `cartpole_swingup` to find one where it started to learn. Note that in the colored MNIST task, they used a penalty weight of 1000, whereas we found no learning to occur until we reduced the penalty weight to 0.01.

Parameter name	Value
Replay buffer capacity	1000000
Batch size	1024
Discount γ	0.99
Optimizer	Adam
Critic learning rate	10^{-5}
Critic target update frequency	2
Critic Q-function soft-update rate τ_{Q}	0.005
Critic encoder soft-update rate τ_{enc}	0.005
Actor learning rate	10^{-5}
Actor update frequency	2
Actor log stddev bounds	$[-5, 2]$
Encoder learning rate	10^{-5}
Decoder learning rate	10^{-5}
Decoder weight decay	10^{-7}
L1 regularization weight	10^{-5}
Temperature learning rate	10^{-4}
Temperature Adam's β_1	0.9
Init temperature	0.1

Table 1. A complete overview of used hyper parameters.

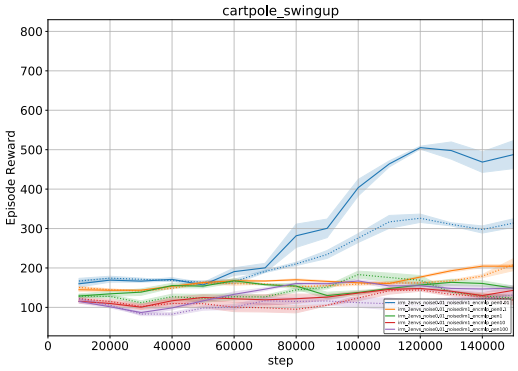

Figure 8. Hyperparameter sweep for IRM on `cartpole_swingup`. All penalty weights fail to learn, even on the training environments, until the penalty weight is very small.

D. Additional Results: Reinforcement Learning

We find that even without noise on the ground truth states, with only two environments, baseline SAC fails as seen in Figure 9.

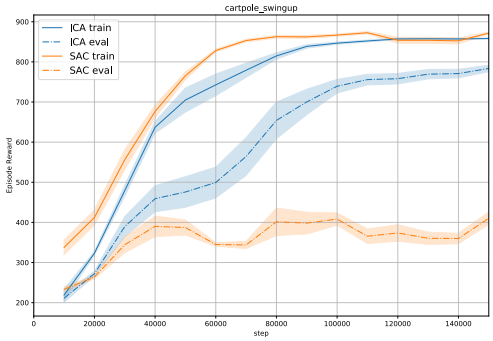

Figure 9. Generalization gap in SAC performance with 2 training environments on Cartpole Swingup from DMC. Evaluated with 10 seeds, standard error shaded.