Domain Adaptive Imitation Learning - Supplementary Materials

A. High-level Comparison to Baselines

Table 2. Comparison of baselines by attributes demonstrated in the paper. The "No Act" column denotes whether or not the demonstrations need to contain actions.

	UNPAIRED DATA	ZEROSHOT IMIT.	EMBOD. MISMATCH	VIEWPOINT MISMATCH	SINGLE-DOMAIN DEMO.	NO ACT.
TPIL (STADIE ET AL., 2017)	✓	✗	✗	✓	✓	✗
IF (GUPTA ET AL., 2017)	✗	✗	✓	✗	✗	✗
IFO (LIU ET AL., 2018)	✗	✗	✗	✓	✗	✓
TCN (SERMANET ET AL., 2018)	✗	✗	✗	✓	✓	✓
GAMA (OURS)	✓	✓	✓	✓	✗	✗

We note that methods such as IF has potential to be applied to the viewpoint mismatch problem and IfO, TCN have the potential to be applied to the embodiment mismatch problem, albeit they were not shown in the paper. TCN has shown mappings between humans and robots can be learned. However they haven't shown that robots can use these mappings to learn from human demonstrations. Below we summarize the key differences between GAMA and the main baselines.

1. We propose an *unsupervised MDP alignment* algorithm (GAMA) capable of learning state correspondences from *unpaired, unaligned demonstrations* while (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018) obtain these correspondences from paired, time-aligned trajectories. Our demonstrations have varying length (up to 2x difference) and diverse starting positions. Since good observation correspondences are prerequisites to the success of (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018), our work provide the missing ingredient. Future work could try learning alignments with GAMA, then apply methods from (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018) to perform CDIL when action information is unavailable from demonstrations.

2. *We remove the need for an expensive RL procedure on a new target task*, by leveraging action information for zero-shot imitation. By learning a composite self policy with both state and action maps, we obtain a near-optimal self policy on new tasks without any environment interactions while prior approaches (Gupta et al., 2017; Liu et al., 2018; Sermanet et al., 2018) require an additional RL step that involves self domain environment interactions.

3. *We use a single algorithm to address both the viewpoint and embodiment mismatch* which have previously been dealt with different solutions.

B. GAMA model architecture

The state, action map $f_{\theta_f}, g_{\theta_g}$, inverse state map $f^{-1}_{\theta_{f^{-1}}}$, transition function $P^x_{\theta_P}$, and discriminators $\{D_{\theta^i_D}\}_{i=1}^N$ are neural networks with hidden layers of size $(200, 200)$. The fitted policies $\{\pi_{y,\mathcal{T}_i}\}_{i=1}^N$ for GAMA-PA and $\pi_{x,\mathcal{T}}$ for GAMA-DA all have hidden layers of size $(300, 200)$. All models are trained with Adam optimizers (Kingma & Ba, 2014) using decay rates $\beta_1 = 0.9, \beta_2 = 0.999$. For the spatial autoencoders used in GAMA-PA-img and GAMA-DA-img, we use the same architecture as in (Finn et al., 2015) We use a learning rate of $1e$-4 for the alignment maps and $1e$-5 for all other components. These parameters are fixed across all experiments.

C. Baseline Implementation Details

In this section we describe our implementation details of the baselines.

Obtaining State Correspondences We use 5000 sampled trajectories in both expert and self domains to learn the state map for IF and CCA. For UMA, we use 20 sampled trajectories to learn that in pendulum and cartpole environment and 50 trajectories in reacher, reacher-tp environment (much beyond these numbers UMA is computationally intractable). For IF and IfO, we use Dynamic Time Warping (DTW) (Muller, 2007) to obtain state correspondences. For IF, DTW uses the (learned) feature space as a metric space to estimate the domain correspondences. For IfO, DTW is applied on the state space. We follow the implementation procedure in (Gupta et al., 2017).

To visualize and quantitatively evaluate the statemaps learned in prior work, we compose the encoder and decoder for IF and use the Moore-Penrose pseudo inverse of the embedding matrix for UMA and CCA.

Transfer Learning In the transfer learning phase for CCA, UMA, IF, and IfO they define a proxy reward function on the target task by using the state correspondence map.

$$r_{\text{proxy}}(s_x^{(t)}) = \frac{1}{|\mathrm{T}|} \sum_{\tau \in \mathrm{T}} \|f(s_{y,\tau}^{(t)}) - g(s_x^{(t)})\|_2^2$$

, where $s_x^{(t)}$ is a self domain state at time t, T is the collection of expert demonstrations, and $s_{y,\tau}^{(t)}$ is the expert domain state at time t in trajectory τ. IfO additionally defines a penalty reward for deviating from states encountered during training. We refer readers to their paper (Liu et al., 2018) for further details. The transferability results of Figure **??** (Right) show the learning curve for training on the ground truth reward for the target task where the policy is pretrained with a training procedure on the proxy reward. All RL steps are performed using the Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015) algorithm.

Architecture For UMA and CCA, the embedding dimension is the minimum state dimension between the expert and self domains. For UMA, we use one state sample every 5 timesteps to reduce the computational time, and we match the pairwise distance matrix of 3-nearest neighbors.

For IF, we use 2 hidden layer with 64 hidden units each and leaky ReLU non-linearities to parameterize embedding function and decoders, the dimension of common feature space is set to be 64. The optimizer are same with respect to our models and the learning rate is $1e$-3. For IfO, we use the same architecture as the statemap in GAMA for their observation conditioned statemap. For TPIL, we use 3 hidden layer with 128 hidden units each and ReLU non-linearities to parameterize the feature extractor, classifier and domain classifier. We use Adam Optimizer with default decay rates and learning rate $1e$-3 to train the discriminator and use same optimizer and learning rate with respect to our model to train the policy.

D. Environments and DAIL tasks

We use the 'Pendulum-v0', 'Cartpole-v0' environments for the pendulum and cartpole tasks which have state space $(w, \dot{w})$ and $(w, \dot{w}, x, \dot{x})$, respectively, where w is the angle of the pendulum/pole and x is the position of the cart. The action spaces are (F_w) and (F_x) where F_w is the torque applied to the pendulum's pivot and F_x is the x-direction force applied to the cart. For snake3, snake4 we use an extension (Wang et al., 2018) of the 'Swimmer-v0' environment from Gym (Brockman et al., 2016). A K link snake has a state representation $(w_1, ..., w_K, \dot{w_1}, ..., \dot{w_K})$ where w_k is the angle of the k^{th} snake joint. The action vector has the form $(F_{w_1}, ...F_{w_K})$ where F_{w_k} is the torque applied to joint k. All reacher environments were extended from the 'Reacher-v0' gym environment. A k link reacher has a state vector of the form $(c_1, ..., c_K, \dot{w_1}, ..., \dot{w_K}, x_g, y_g)$ where c_k is the coordinates of the k^{th} reacher joint and (x_g, y_g) is the position of the goal. Note the key difference with the original Reacher-v0 environment is that we use coordinates of joints instead of joint angles and the difference vector between the end effector and the goal coordinate was removed from the state to make the task more challenging. The action vector has the form $(F_{w_1}, ...F_{w_K})$ where F_{w_k} is the torque applied to joint k. Below we specifically describe each DAIL task. The statemap acts only on the non-goal dimensions.

Dynamics-Reach2Reach (D-R2R): Self domain is reach2 and expert domain is reach2 with isotropic gaussian noise injected into the dynamics. State, action spaces are the same the one for a k-link reacher with $k = 2$. The N alignment tasks are reaching for N goals near the wall of the arena and the target tasks are reaching for 12 new goals near the corner of the arena. The new goals are placed as far as possible from the alignment task goals within the bounds of the arena to make the task more challenging.

Dynamics-Reach2Push (D-R2P): Same as D-R2R except the target task is pushing a block to a goal location. State, action spaces are the same the one for a k-link reacher with $k = 2$. Here the goal location represents the location to push the block to. Block is always initialized in the same location.

Embodiment-Reach2Reach (E-R2R): Self domain is reach2 and expert is reach3. Rest is the same as D-R2R.

Embodiment-Reach2Push (E-R2P): Self domain is reach2 and expert is reach3. Rest is the same as D-R2P.

Viewpoint-Reach2Reach (V-R2R): Self domain is reach2 and expert domain is reach2-tp1 that has the same "third person" view state space as that in (Stadie et al., 2017) with a 30° planar offset. The state space is a projection of the joint coordinates onto the offset viewing plane, e.g a joint coordinate $(1, 1)$ in the self domain is corresponds to $(1, 0.7)$ in the expert domain. The alignment/target tasks are the same as D-R2R.

Viewpoint-Reach2Write (V-R2W): Self domain is reach2 and expert domain is reach2-tp2 that has a different "third person" view state space with a $180°$ axial offset. Thus a joint coordinate $(1, 1)$ in the self domain corresponds to $(-1, -1)$ in the expert domain. We use the robot's joint level state-action space. The N alignment tasks are reaching for N goals and the target task is tracing letters as fast as possible. The goal location in the writing task represents the next vertex of the letter to trace. Once the first vertex is reached, the goal coordinates are updated to be the next vertex coordinates. The reward function is defind as follows:

$$R_{write}(s) = \begin{cases} 100 & \text{if state } s \text{ corresponds to reaching a vertex} \\ -1 & \text{else} \end{cases}$$

Thus the agent must perform a sequential reaching task and accomplish it as fast as possible. The key difference with a normal reaching task is that the reacher must not slow down at each vertex and plan it's path accordingly in order to minimize drastic direction changes. Further more the reward is significantly more sparse than the original reacher reward which gets reward inversely proportional to the distance between the end effector and the goal.

E. MDP Alignment Visualization

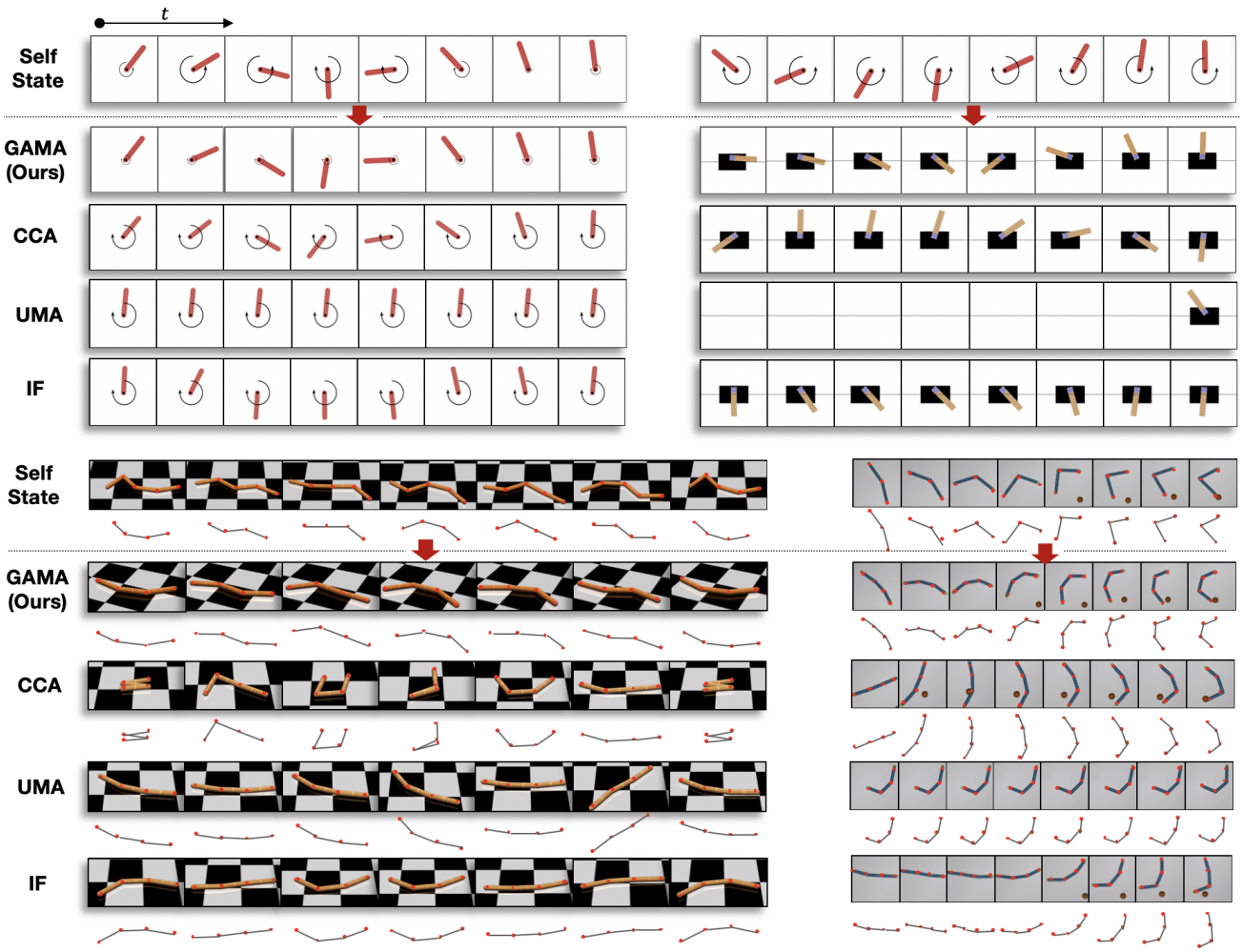

Figure 7. **MDP Alignment Visualization (Extended)**. The state maps learned by GAMA and baselines are shown for pen↔pen (Top Left), pen↔cart (Top Right), snake4↔snake3 (Bottom Left), reach2↔reach3 (Bottom Right). See Appendix E to see more baselines. GAMA is able to recover MDP permutations for alignable pairs pen↔pen, pen↔cart and find meaningful correspondences between "weakly alignable" pairs snake4↔snake3, reach2↔reach3. For pen↔cart, UMA learns a statemap that outputs out-of-bounds coordinates mainly because the pendulum demonstrations are concentrated around the pole upright state. The optimal UMA embedding matrix in this case is a zero matrix. Then the UMA state map matrix norm is proportional to the inverse embedding matrix norm which is very large. See https://youtu.be/l0tc1JCN_1M for videos

F. Proofs

We start by introducing definitions and assumptions that will be used in proving both Theorem 1, 2

Definition 5. *An optimal policy π_x is **covering** if $O_{\mathcal{M}_x}(s_x, a_x) = 1 \Rightarrow a_x \in supp(\pi_x(\cdot|s_x))$.*

Definition 6. *MDP $\mathcal{M}_x$ is **unichain**, if all policies induce irreducible Markov Chains and all stochastic optimal policies induce ergodic, i.e irreducible and aperiodic, Markov Chains.*

Assumption 1. *All considered MDPs are unichain with discrete state, action spaces and deterministic dynamics i.e. $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. Furthermore, there exists dummy state, actions s^d, a^d where $O_{\mathcal{M}}(s, a^d) = 0 \; \forall s \in \mathcal{S}$ and $O_{\mathcal{M}}(s^d, a) = 0 \; \forall a \in \mathcal{A}$*

As stated in Assumption 1, we consider discrete unichain MDPs with deterministic dynamics. This assumption is weak since physics is largely deterministic and many control behaviors, such as walking, are described by unichains.

F.1. Proof of Theorem 1

Theorem 1. *Let $\mathcal{M}_x, \mathcal{M}_y$ be MDPs satisfying Assumption 1 (see Appendix F), $\mathcal{M}_x \geq_{\phi,\psi} \mathcal{M}_y$, and π_y be optimal in $\mathcal{M}_y$. Then, $\forall g : \mathcal{A}_y \to \mathcal{A}_x$ s.t $\psi \circ g(a_y) = a_y \;\; \forall a_y \in \{a_y | \exists s_y \in \mathcal{S}_y \; s.t \; O_{\mathcal{M}_y}(s_y, a_y) = 1\}$, it holds that $\hat{\pi}_x = g \circ \pi_y \circ \phi$ is optimal in $\mathcal{M}_x$.*

Proof. Without loss of generality, consider an arbitrarily chosen sample $a_x = g(a_y), a_y \sim \pi_y(\cdot|\phi(s_x))$ for any $s_x \in \mathcal{S}_x$. We first see that:

$$O_{\mathcal{M}_y}\big(\phi(s_x), \psi(a_x)\big) = O_{\mathcal{M}_y}\big(\phi(s_x), \psi(g(a_y))\big) = O_{\mathcal{M}_y}\big(\phi(s_x), a_y\big) = 1 \tag{9}$$

where the first step substitutes $a_x = g(a_y)$, the second step applies $\psi \circ g(a_y) = a_y$ since $O(\phi(s_x), a_y) = 1$ due to the optimality of π_y, and the last step follows from Corollary 1. Since (ϕ, ψ) is a reduction, we have that $O_{\mathcal{M}_y}(\phi(s_x), \psi(a_x)) = 1 \Rightarrow O_{\mathcal{M}_x}(s_x, a_x) = 1$ by Equation (1). Therefore, $O_{\mathcal{M}_x}(s_x, a_x) = 1 \;\; \forall s_x \in \mathcal{S}_x, \forall a_x \in \text{supp}(\hat{\pi}_x(\cdot|s_x))$. Then by Lemma 1, $\hat{\pi}_x$ is optimal. □

F.2. Proof of Theorem 2

We first introduce some lemmas necessary to proving our main theorem.

Lemma 1. *Let MDP $\mathcal{M}_x$ satisfy Assumption 1 and $\pi_x(a_x|s_x)$ be a (stochastic) mixture policy that chooses a_x randomly from $\{a_x | O(s_x, a_x) = 1\}$. Then, π_x is optimal. (Ortner, 2005)*

Corollary 1. *Let MDP $\mathcal{M}_x$ satisfy Assumption 1 and π_x be optimal. Then $O_{\mathcal{M}_x}(s_x, a_x) = 1 \; \forall s_x \in \mathcal{S}_x, a_x \in supp(\pi_x(\cdot|s_x))$*

Lemma 2. *Let MDP $\mathcal{M}_x$ satisfy Assumption 1 and π_x be a stochastic optimal policy. Then the triplet stationary distribution $\rho^x_{\pi_x}(s_x, a_x, s'_x) = \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x, s_x^{(t+1)} = s'_x; \pi_x, P_x, \eta_x)$ exists and is unique.*

Proof.

$$\begin{aligned}
\rho^x_{\pi_x}(s_x, a_x, s'_x) &= \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x, s_x^{(t+1)} = s'_x; \pi_x, P_x, \eta_x) \\
&= \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x; \pi_x, P_x, \eta_x) \pi_x(a_x|s_x) \mathbb{1}(s'_x = P_x(s_x, a_x)) \\
&= \pi_x(a_x|s_x) \mathbb{1}(s'_x = P_x(s_x, a_x)) \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x; \pi_x, P_x, \eta_y)
\end{aligned}$$

where $\mathbb{1}$ is the indicator function. The limit in the last line is the stationary distribution over states, which exists and is unique since a stochastic optimal policy induces an ergodic Markov Chain over states. □

Lemma 3. *If a real sequence $\{a_i\}_{i=1}^{\infty}$ converges to some $a \in \mathbb{R}$, then*

$$\lim_{T\to\infty} \frac{1}{T} \textstyle\sum_{i=1}^{T} a_i = \lim_{i\to\infty} a_i = a$$

Proof. Denote $A_T = \sum_{i=1}^{T} a_i$, and $B_T = T$. We have

$$\lim_{T\to\infty} \frac{A_{T+1} - A_T}{B_{T+1} - B_T} = \lim_{T\to\infty} a_{T+1} = a \tag{10}$$

According to the Stolz–Cesàro theorem,

$$\lim_{T\to\infty} \frac{A_{T+1} - A_T}{B_{T+1} - B_T} = \lim_{T\to\infty} \frac{A_T}{B_T}$$

if the limit on the left hand side exists. Therefore

$$\lim_{T\to\infty} \frac{A_T}{B_T} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} a_i = a \tag{11}$$

which completes the proof. □

Recall that our target distribution $\sigma^y_{\pi_y}$ and proxy distribution $\sigma^{x\to y}_{\hat{\pi}_x}$ were defined as:

$$\sigma^y_{\pi_y}(s_y, a_y, s'_y) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(s_y^{(t)} = s_y, a_y^{(t)} = a_y, s_y^{(t+1)} = s'_y; \pi_y, P_y, \eta_y) \tag{12}$$

$$\sigma^{x\to y}_{\hat{\pi}_x}(s_y, a_y, s'_y) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(\hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y, \hat{s}_y^{(t+1)} = s'_y; \mathcal{P}) \tag{13}$$

We are now ready to prove that our proxy and target limiting distributions exist.

Lemma 4. *Let MDP $\mathcal{M}_y$ satisfy Assumption 1 and π_y be a stocahstic optimal policy. Then, $\sigma^y_{\pi_y}(s_y, a_y, s'_y) = \rho^y_{\pi_y}(s_y, a_y, s'_y)$.*

Proof. Recall that the stationary distribution $\rho^y_{\pi_y}(s_y, a_y, s'_y)$ is the following limiting distribution:

$$\rho^y_{\pi_y}(s_y, a_y, s'_y) = \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y, a_y^{(t)} = a_y, s_y^{(t+1)} = s'_y; \pi_y, P_y, \eta_y) \tag{14}$$

$\rho^y_{\pi_y}(s_y, a_y, s'_y)$ exist for $\mathcal{M}_y$ as shown in Lemma 2. Then,

$$\begin{aligned}
\sigma^y_{\pi_y}(s_y, a_y, s'_y) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \Pr(s_y^{(t)} = s_y, a_y^{(t)} = a_y, s_y^{(t+1)} = s'_y; \pi_y, P_y, \eta_y) && (15)\\
&= \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y, a_y^{(t)} = a_y, s_y^{(t+1)} = s'_y; \pi_y, P_y, \eta_y) && (16)\\
&= \rho^y_{\pi_y}(s_y, a_y, s'_y) && (17)
\end{aligned}$$

as desired. The second line follows from Lemma 3 and the last line follows from Lemma 2. □

Lemma 5. *Let MDP $\mathcal{M}_y$ satisfy Assumption 1 and π_y be a stochastic optimal policy. Then,*

$$supp(\sigma^y_{\pi_y}) \subseteq \{(s_y, a_y, s'_y) | O_{\mathcal{M}_y}(s_y, a_y) = 1, s_y, s'_y \in \mathcal{S}_y, a_y \in \mathcal{A}_y\}$$

Proof. Assume for contradiction that there exists $(s_y, a_y, s'_y) \in \text{supp}(\sigma^y_{\pi_y})$ but $(s_y, a_y, s'_y) \notin \{(s_y, a_y, s'_y) | O_{\mathcal{M}_y}(s_y, a_y) = 1, s_y, s'_y \in \mathcal{S}_y, a_y \in \mathcal{A}_y\}$. Then $O_{\mathcal{M}_y}(s_y, a_y) = 0$. Since

$$\begin{aligned}
\sigma^y_{\pi_y}(s_y, a_y, a_y) &= \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y, a_y^{(t)} = a_y, s_y^{(t+1)} = s'_y; \pi_y, P_y, \eta_y) \\
&= \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \Pr(a_y^{(t)} = a_y | s_y^{(t)} = s_y) \Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \\
&= \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \pi_y(a_y|s_y) \Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \\
&= \cancelto{0}{\pi_y(a_y|s_y)} \Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \\
&= 0
\end{aligned}$$

First line follows from Lemma 4 and terms are taken out of the limit in the second to last line since the stationary distribution over states exist as $\mathcal{M}_y$ is unichain and π_y is stochastic optimal. $\pi_y(a_y|s_y) = 0$ since $O_{\mathcal{M}_y}(s_y, a_y) = 0 \Rightarrow \pi_y(a_y|s_y) = 0$ from Corollary 1. Then, we have $\sigma^y_{\pi_y}(s_y, a_y, a_y) = 0$ which contradicts $(s_y, a_y, s'_y) \in \text{supp}(\sigma^y_{\pi_y})$ concluding the proof. □

Lemma 6. *Let MDP $\mathcal{M}_x$ satisfy Assumption 1 and $\hat{\pi}_x = g \circ \pi_y \circ f$ be an stochastic optimal policy in $\mathcal{M}_x$ where $f : \mathcal{S}_x \to \mathcal{S}_y$ is the state map, $g : \mathcal{A}_y \to \mathcal{A}_x$ is injective action map, and π_y is a stochastic optimal policy in $\mathcal{M}_y$. Further let $\mathcal{F} : \mathcal{S}_x \times g(\mathcal{A}_y) \times \mathcal{S}_x \to \mathcal{S}_y \times \mathcal{A}_y \times \mathcal{S}_y$ be the map $\mathcal{F}(a, b, c) = (f(a), g^{-1}(b), f(c))$. Then, $\sigma^{x\to y}_{\hat{\pi}_x}(s_y, a_y, s'_y) = \mathcal{F}(\rho^x_{\hat{\pi}_x}(s_x, a_x, s'_x))$.*

Proof. We first define the triplet random variables $X^{(t)} = (s_x^{(t)}, a_x^{(t)}, s_x^{(t+1)})$ for $t = 0, 1, 2, ...$ where $s_x^{(t)}, a_x^{(t)}, s_x^{(t+1)}$ for $t = 0, 1, 2, ...$ were defined in Definition 4. $\mathcal{F}$ is a function on $supp(\rho^x_{\hat{\pi}_x}) \in \mathcal{S}_x \times g(\mathcal{A}_y) \times \mathcal{S}_x$ and $\mathcal{F}(X^{(t)}) = (\hat{s}_y^{(t)}, \hat{a}_y^{(t)}, \hat{s}_y^{(t+1)})$. Furthermore, since $\mathcal{F}$ is a function defined on a discrete domain and codomain, there always exists a trivial continuous extension of $\mathcal{F}$. We may thus apply the continuous mapping theorem (Billingsley, 1968):

$$X^{(t)} \xrightarrow{d} X \Rightarrow \mathcal{F}(X^{(t)}) \xrightarrow{d} \mathcal{F}(X)$$

Since $\mathcal{M}_x$ is unichain and $\hat{\pi}_x$ is stochastic optimal, the distribution of $X^{(t)}$ converges (in distribution) to $\rho^x_{\hat{\pi}_x}(s_x, a_x, s'_x)$ as $t \to \infty$ by Lemma 2. Applying the continuous mapping theorem, it follows that the distribution of $\mathcal{F}(X^{(t)}) = (\hat{s}_y^{(t)}, \hat{a}_y^{(t)}, \hat{s}_y^{(t+1)})$ converges (in distribution) to the pushforward measure $\mathcal{F}(\rho^x_{\hat{\pi}_x}(s_x, a_x, s'_x))$ as $t \to \infty$

Directly applying this result, we obtain:

$$\begin{aligned}
\sigma^{x\to y}_{\hat{\pi}_x}(s_y, a_y, s'_y) &= \lim_{T\to\infty} \frac{1}{T} \textstyle\sum_{t=0}^{T-1} \Pr(\hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y, \hat{s}_y^{(t+1)} = s'_y; \mathcal{P}) && (18) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y, \hat{s}_y^{(t+1)} = s'_y; \mathcal{P}) && (19) \\
&= \mathcal{F}(\rho^x_{\hat{\pi}_x}(s_x, a_x, s'_x)) && (20)
\end{aligned}$$

as desired. Line (18) → (19) follows from Lemma 3 and (19) → (20) follows from the continuous mapping theorem. □

Lemma 7. *Let X, Y be countable sets, $\phi : X \to Y$ be a function, and $\mathbb{1}$ be the indicator function. We denote $\phi^{-1}(y) = \{x | \phi(x) = y\}$. Then $\forall x \in X, y \in Y$*

$$\mathbb{1}(y = \phi(x)) = \textstyle\sum_{z\in\phi^{-1}(y)} \mathbb{1}(x = z)$$

Proof. Since both the left and right hand-side of the desired equality only take on values in $\{0, 1\}$, it suffices to show the following statements hold for arbitrarily chosen $x \in X, y \in Y$:

$$\begin{aligned}
\textstyle\sum_{z\in\phi^{-1}(y)} \mathbb{1}(x = z) = 1 &\Rightarrow \mathbb{1}(y = \phi(x)) = 1 \\
\mathbb{1}(y = \phi(x)) = 1 &\Rightarrow \textstyle\sum_{z\in\phi^{-1}(y)} \mathbb{1}(x = z) = 1
\end{aligned}$$

For the first direction, we see that if $\sum_{z\in\phi^{-1}(y)}\mathbb{1}(x=z)=1$, then $x\in\phi^{-1}(y)$, and thus $\phi(x)=y$.

For the second direction if $\mathbb{1}(y=\phi(x))=1$, then $x\in\phi^{-1}(y)$. Thus there exists a unique z such that $z=x$ and $z\in\phi^{-1}(y)$. Then, $\sum_{z\in\phi^{-1}(y)}\mathbb{1}(x=z)=1$ as desired, which concludes the proof. □

Finally, we prove the main theorem. Recall that the optimization objectives are: 1. optimality of $\hat{\pi}_x$ 2. $\sigma^{x\to y}_{\hat{\pi}_x}=\sigma^{y}_{\pi_y}$.

Theorem 2. *Let $\mathcal{M}_x,\mathcal{M}_y$ be MDPs satisfying Assumption 1 (see Supp Materials). If $\mathcal{M}_x\geq\mathcal{M}_y$, then $\exists f:\mathcal{S}_x\to\mathcal{S}_y, g:\mathcal{A}_y\to\mathcal{A}_x$, and an optimal covering policy π_y (see Appendix F) that satisfy objectives 1 and 2. Conversely, if $\exists f:\mathcal{S}_x\to\mathcal{S}_y$, an injective map $g:\mathcal{A}_y\to\mathcal{A}_x$ and an optimal covering policy π_y satisfying objectives 1 and 2, then $\mathcal{M}_x\geq\mathcal{M}_y$ and $\exists(\phi,\psi)\in\Gamma(\mathcal{M}_x,\mathcal{M}_y)$ s.t $f=\phi$ and $\psi\circ g(a_y)=a_y,\forall a_y\in\mathcal{A}_y$.*

Proof. We first show the ($\Rightarrow$) direction. Using any $(\phi,\psi)\in\Gamma(\mathcal{M}_x,\mathcal{M}_y)$ we construct f and g in the following manner: $f(s_x)=\phi(s_x)\ \ \forall s_x\in\mathcal{S}_x$. $g(a_y)$ maps to an arbitrary chosen element from the set $\psi^{-1}(a_y)=\{a_x|\psi(a_x)=a_y\}$ if $\psi^{-1}(a_y)\neq\emptyset$ and an arbitrarily chosen action $a_x\in\mathcal{A}_x$ otherwise. We see that $\forall a_y\in\mathcal{A}_y$ for which $\exists s_y\in\mathcal{S}_y$ such that $O_{\mathcal{M}_y}(s_y,a_y)=1$, it holds that $\psi^{-1}(a_y)\neq\emptyset$ by Eq 2. Therefore, $\psi\circ g(a_y)=a_y\ \ \forall a_y\in\mathcal{A}_y$ for which $\exists s_y$ such that $O_{\mathcal{M}_y}(s_y,a_y)=1$ since ψ maps all elements in $\psi^{-1}(a_y)$ to a_y. For π_y we choose any covering optimal policy for $\mathcal{M}_y$. It suffices to show that this choice of f,g,π_y satisfies objectives 1, 2.

• Objective 1. $\hat{\pi}_x$ is optimal: follows from Lemma 1.

• Objective 2. $\sigma^{x\to y}_{\hat{\pi}_x}=\sigma^{y}_{\pi_y}$: Since $f=\phi$ is a reduction, it follows that $\forall s_y\in\mathcal{S}_y,a_y\in\mathcal{A}_y$ such that $O_{\mathcal{M}_y}(s_y,a_y)=1$, any $s'_y\in\mathcal{S}_y$, and $\forall t=0,1,2,...$:

$$
\begin{aligned}
&\Pr(\hat{s}_y^{(t+1)}{=}s'_y|\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\\
&=\sum_{s'_x\in\mathcal{S}_x}\Pr(\hat{s}_y^{(t+1)}{=}s'_y|\hat{s}_x^{(t+1)}{=}s'_x,\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\Pr(\hat{s}_x^{(t+1)}{=}s'_x|\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\\
&=\sum_{s'_x\in\mathcal{S}_x}\Pr(\hat{s}_y^{(t+1)}{=}s'_y|\hat{s}_x^{(t+1)}{=}s'_x)\sum_{\substack{s_x\in\mathcal{S}_x\\a_x\in\mathcal{A}_x}}\Pr(\hat{s}_x^{(t+1)}{=}s'_x|s_x^{(t)}{=}s_x,a_x^{(t)}{=}a_x,\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\Pr(s_x^{(t)}{=}s_x,a_x^{(t)}{=}a_x|\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\\
&=\sum_{s'_x\in\mathcal{S}_x}\mathbb{1}(s'_y{=}\phi(s'_x))\sum_{\substack{s_x\in\mathcal{S}_x\\a_x\in\mathcal{A}_x}}\Pr(\hat{s}_x^{(t+1)}{=}s'_x|s_x^{(t)}{=}s_x,a_x^{(t)}{=}a_x)\Pr(s_x^{(t)}{=}s_x|a_x^{(t)}{=}a_x,\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\Pr(a_x^{(t)}{=}a_x|\hat{s}_y^{(t)}{=}s_y,\hat{a}_y^{(t)}{=}a_y)\\
&=\sum_{s'_x\in\mathcal{S}_x}\mathbb{1}(s'_y{=}\phi(s'_x))\sum_{\substack{s_x\in\mathcal{S}_x\\a_x\in\mathcal{A}_x}}\mathbb{1}(s'_x{=}P_x(s_x,a_x))\Pr(s_x^{(t)}{=}s_x|\hat{s}_y^{(t)}{=}s_y)\Pr(a_x^{(t)}{=}a_x|\hat{a}_y^{(t)}{=}a_y)\\
&=\sum_{s'_x\in\mathcal{S}_x}\mathbb{1}(s'_y{=}\phi(s'_x))\sum_{\substack{s_x\in\mathcal{S}_x\\a_x\in\mathcal{A}_x}}\mathbb{1}(s'_x{=}P_x(s_x,a_x))\frac{\Pr(\hat{s}_y^{(t)}{=}s_y|s_x^{(t)}{=}s_x)\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\mathcal{S}_x}\Pr(\hat{s}_y^{(t)}{=}s_y|s_x^{(t)}{=}s''_x)\Pr(s_x^{(t)}{=}s''_x)}\mathbb{1}(a_x{=}g(a_y))\\
&=\sum_{s'_x\in\mathcal{S}_x}\mathbb{1}(s'_y{=}\phi(s'_x))\sum_{s_x\in\mathcal{S}_x}\mathbb{1}(s'_x{=}P_x(s_x,g(a_y)))\frac{\mathbb{1}(s_y{=}\phi(s_x))\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\mathcal{S}_x}\mathbb{1}(s_y{=}\phi(s''_x))\Pr(s_x^{(t)}{=}s''_x)}\\
&=\sum_{s'_x\in\phi^{-1}(s'_y)}\sum_{s_x\in\phi^{-1}(s_y)}\mathbb{1}(s'_x{=}P_x(s_x,g(a_y)))\frac{\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\phi^{-1}(s_y)}\Pr(s_x^{(t)}{=}s''_x)}\\
&=\sum_{s_x\in\phi^{-1}(s_y)}\frac{\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\phi^{-1}(s_y)}\Pr(s_x^{(t)}{=}s''_x)}\sum_{s'_x\in\phi^{-1}(s'_y)}\mathbb{1}(s'_x{=}P_x(s_x,g(a_y)))\\
&\overset{\text{Lemma7}}{=}\sum_{s_x\in\phi^{-1}(s_y)}\frac{\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\phi^{-1}(s_y)}\Pr(s_x^{(t)}{=}s''_x)}\mathbb{1}\Big(s'_y{=}\phi\Big(P_x\big(s_x,g(a_y)\big)\Big)\Big)\\
&\overset{\text{Eq3}}{=}\sum_{s_x\in\phi^{-1}(s_y)}\frac{\Pr(s_x^{(t)}{=}s_x)}{\sum_{s''_x\in\phi^{-1}(s_y)}\Pr(s_x^{(t)}{=}s''_x)}\mathbb{1}(s'_y{=}P_y(s_y,a_y))\\
&{=}\mathbb{1}(s'_y{=}P_y(s_y,a_y)){=}\ \Pr(s_y^{(t+1)}{=}s'_y|s_y^{(t)}{=}s_y,a_y^{(t)}{=}a_y)
\end{aligned}
\tag{21}
$$

Furthermore, from Definition 4, we have:

$$\Pr(\hat{a}_y^{(t)} = a_y | \hat{s}_y^{(t)} = s_y) = \pi_y(a_y|s_y) = \Pr(a_y^{(t)} = a_y | s_y^{(t)} = s_y) \tag{22}$$

Then, $\forall s_y, s'_y \in \mathcal{S}_y$ and $\forall t = 0, 1, 2, ...$

$$\begin{aligned}
\Pr(\hat{s}_y^{(t+1)} = s_y | \hat{s}_y^{(t)} = s_y) &= \sum_{a_y \in \mathcal{A}_y} \Pr(\hat{s}_y^{(t+1)} = s_y | \hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y) \Pr(\hat{a}_y^{(t)} = a_y | \hat{s}_y^{(t)} = s_y) \\
&= \sum_{a_y \in \text{supp}(\pi_y(\cdot|s_y))} \Pr(s_y^{(t+1)} = s_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \pi_y(a_y|s_y) \\
&= \Pr(s_y^{(t+1)} = s_y | s_y^{(t)} = s_y)
\end{aligned} \tag{23}$$

we are justified in the substitution for the dynamics in the second line since $O_{\mathcal{M}_y}(s_y, a_y) = 1 \forall s_y \in \mathcal{S}_y, a_y \in \text{supp}(\pi_y(\cdot|s_y))$ by Corollary 1. Since $\mathcal{M}_y$ is unichain and π_y is a stochastic optimal policy, the stationary distribution $\lim_{t\to\infty} \Pr(s_y^{(t)} = s_y)$ is invariant to the initial state distribution η_y and only depends on the state transition dynamics $\Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y)$. Equivalently any stochastic process with the same state transition dynamics will converge to the same stationary distribution regardless of the initial state distribution. Thus,

$$\lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y) = \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \quad \forall s_y \in \mathcal{S}_y \tag{24}$$

Finally putting these results together, the following equalities hold for $(s_y, a_y, s'_y) \in \{(s_y, a_y, s'_y) | O_{\mathcal{M}_y}(s_y, a_y) = 1, s_y, s'_y \in \mathcal{S}_y, a_y \in \mathcal{A}_y\}$

$$\begin{aligned}
\sigma^{x\to y}_{\hat{\pi}_x}(s_y, a_y, s'_y) &\overset{\text{Lemma 4}}{=} \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y, \hat{s}_y^{(t+1)} = s'_y; \mathcal{P}) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y) \Pr(\hat{a}_y^{(t)} = a_y | \hat{s}_y^{(t)} = s_y) \Pr(\hat{s}_y^{(t+1)} = s'_y | \hat{s}_y^{(t)} = s_y, \hat{a}_y^{(t)} = a_y) \\
&\overset{\text{Eq (21),(22)}}{=} \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y) \Pr(a_y^{(t)} = a_y | s_y^{(t)} = s_y) \Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \\
&= \pi_y(a_y|s_y) \mathbb{1}(s'_y = P_y(s_y, a_y)) \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y) \\
&\overset{\text{Eq (24)}}{=} \pi_y(a_y|s_y) \mathbb{1}(s'_y = P_y(s_y, a_y)) \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \\
&= \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y) \Pr(a_y^{(t)} = a_y | s_y^{(t)} = s_y) \Pr(s_y^{(t+1)} = s'_y | s_y^{(t)} = s_y, a_y^{(t)} = a_y) \\
&\overset{\text{Lemma 4}}{=} \sigma^y_{\pi_y}(s_y, a_y, s'_y)
\end{aligned}$$

The constant terms are moved in and out of the limit since the stationary distribution over states exist as $\mathcal{M}_y$ is unichain and π_y is optimal in $\mathcal{M}_y$. This allows us to conclude that $\sigma^{x\to y}_{\hat{\pi}_x} = \sigma^y_{\pi_y}$ since $\sigma^y_{\pi_y}$ is supported on $\{(s_y, a_y, s'_y) | O_{\mathcal{M}_y}(s_y, a_y) = 1, s_y, s'_y \in \mathcal{S}_y, a_y \in \mathcal{A}_y\}$ by Lemma 5.

Now we show the ($\Leftarrow$) direction. We first introduce some overloaded notation:

$$\begin{aligned}
\sigma^x_{\hat{\pi}_x}(s_x) &= \lim_{T\to\infty} \frac{1}{T} \textstyle\sum_{t=0}^{T-1} \Pr(s_x^{(t)} = s_x; \hat{\pi}_x, P_x, \eta_x) \overset{\text{Lemma4}}{=} \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x; \hat{\pi}_x, P_x, \eta_x) \\
\sigma^x_{\hat{\pi}_x}(s_x, a_x) &= \lim_{T\to\infty} \frac{1}{T} \textstyle\sum_{t=0}^{T-1} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x; \hat{\pi}_x, P_x, \eta_x) \\
&\overset{\text{Lemma 4}}{=} \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x; \hat{\pi}_x, P_x, \eta_x) \\
&= \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x; \hat{\pi}_x, P_x, \eta_x) \hat{\pi}_x(a_x|s_x) \\
&= \hat{\pi}_x(a_x|s_x) \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x; \hat{\pi}_x, P_x, \eta_x) \\
&= \hat{\pi}_x(a_x|s_x) \sigma^x_{\hat{\pi}_x}(s_x)
\end{aligned} \tag{25}$$

Then,

$$\begin{aligned}
\sigma^x_{\hat{\pi}_x}(s_x, a_x, s'_x) &\overset{\text{Lemma 4}}{=} \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x, s_x^{(t+1)} = s'_x; \hat{\pi}_x, P_x, \eta_x) \\
&= \lim_{t\to\infty} \Pr(s_x^{(t+1)} = s'_x | s_x^{(t)} = s_x, a_x^{(t)} = a_x) \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x; \hat{\pi}_x, P_x, \eta_x) \\
&= \mathbb{1}(s'_x = P_x(s_x, a_x)) \lim_{t\to\infty} \Pr(s_x^{(t)} = s_x, a_x^{(t)} = a_x; \hat{\pi}_x, P_x, \eta_x) \\
&= \mathbb{1}(s'_x = P_x(s_x, a_x)) \sigma^x_{\hat{\pi}_x}(s_x, a_x) \qquad (26) \\
&= \mathbb{1}(s'_x = P_x(s_x, a_x)) \hat{\pi}_x(a_x|s_x) \sigma^x_{\hat{\pi}_x}(s_x) \qquad (27)
\end{aligned}$$

Given f, g that satisfy objective 1, 2, and 3 we construct (ϕ, ψ) as follows and show that $(\phi, \psi) \in \Gamma(\mathcal{M}_x, \mathcal{M}_y)$:

$$\phi(s_x) = \begin{cases} f(s_x) & \text{if } s_x \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x)) \\ s^d_y & \text{otherwise} \end{cases}$$

$$\psi(a_x) = \begin{cases} g^{-1}(a_x) & \text{if } a_x \in \mathcal{A}_{\hat{\pi}_x} = \bigcup_{s_x \in \mathcal{S}_x} \text{supp}(\hat{\pi}_x(\cdot|s_x)) \\ a^d_y & \text{otherwise} \end{cases}$$

where s^d_y, a^d_y are dummy state, actions such that $O_{\mathcal{M}_y}(s^d_y, a_y) = 0 \ \ \forall a_y \in \mathcal{A}_y$ and $O_{\mathcal{M}_y}(s_y, a^d_y) = 0 \ \ \forall s_y \in \mathcal{S}_y$. Such dummy state, actions always exist per Assumption 1. Mapping to the dummy state, action will ensure that the constructions will not map suboptimal state, action pairs from domain x to optimal state action pairs in domain y. The following statement holds for our construction (ϕ, ψ):

$$(s^*_x, a^*_x) \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x)) \iff O_{\mathcal{M}_y}(\phi(s^*_x), \psi(a^*_x)) = 1 \quad \forall s^*_x \in \mathcal{S}_x, a^*_x \in \mathcal{A}_x \qquad (28)$$

We first prove the forward direction: $(s^*_x, a^*_x) \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x)) \Rightarrow \sigma^x_{\hat{\pi}_x}(s^*_x, a^*_x) \overset{\text{Eq25}}{=} \sigma^x_{\hat{\pi}_x}(s_x)\hat{\pi}_x(a^*_x|s^*_x) > 0$, so $\sigma^x_{\hat{\pi}_x}(s^*_x) > 0$, i.e $s^*_x \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x))$, and $\hat{\pi}_x(a^*_x|s^*_x) > 0$. Furthermore, $\hat{\pi}_x(a^*_x|s^*_x) > 0 \Rightarrow g^{-1}(a^*_x) \in \text{supp}(\pi_y(\cdot|f(s^*_x)))$ since g is injective. To see this, assume $\exists (s^*_x, a^*_x)$ such that $\hat{\pi}_x(a^*_x|s^*_x) > 0$ but $g^{-1}(a^*_x) \notin \text{supp}(\pi_y(\cdot|f(s^*_x)))$. Then there must exists $a'_y \in \text{supp}(\pi_y(\cdot|f(s^*_x)))$ such that $a'_y \neq g^{-1}(a^*_x)$ but $g(a'_y) = g(g^{-1}(a^*_x)) = a^*_x$ contradicting the injectivity of g on $\mathcal{A}_y$. Putting these results together we obtain $\psi(a^*_x) = g^{-1}(a^*_x) \in \text{supp}(\pi_y(\cdot|\phi(s^*_x)))$. Since π_y is a stochastic optimal policy and $\mathcal{M}_y$ is unichain, $\psi(a^*_x) \in \text{supp}(\pi_y(\cdot|\phi(s^*_x))) \Rightarrow O_{\mathcal{M}_y}(\phi(s^*_x), \psi(a^*_x)) = 1$ by Corollary 1.

For the converse direction we prove the contrapostive: $(s^*_x, a^*_x) \notin \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x)) \Rightarrow O_{\mathcal{M}_y}(\phi(s^*_x), \psi(a^*_x)) = 0 \ \ \forall s_x \in \mathcal{S}_x, a_x \in \mathcal{A}_x$. We exhaustively consider all cases in which $(s^*_x, a^*_x) \notin \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x))$, i.e $\sigma^x_{\hat{\pi}_x}(s^*_x, a^*_x) \overset{\text{Eq25}}{=} \hat{\pi}_x(a^*_x|s^*_x)\sigma^x_{\hat{\pi}_x}(s^*_x) = 0$. If $\sigma^x_{\hat{\pi}_x}(s^*_x) = 0$, then $s^*_x \notin \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x))$, so $O_{\mathcal{M}_y}(\phi(s^*_x), a_y) = O_{\mathcal{M}_y}(s^d_y, a_y) = 0 \ \ \forall a_y \in \mathcal{A}_y$. Else if $\hat{\pi}_x(a^*_x|s^*_x) = 0, \sigma^x_{\hat{\pi}_x}(s^*_x) > 0$ and $a_x \notin \mathcal{A}_{\hat{\pi}_x}$ then $O_{\mathcal{M}_y}(s_y, \psi(a^*_x)) = O_{\mathcal{M}_y}(s_y, a^d_y) = 0 \ \ \forall s_y \in \mathcal{S}_y$. Finally, consider the case $\hat{\pi}_x(a^*_x|s^*_x) = 0, \sigma^x_{\hat{\pi}_x}(s^*_x) > 0$ and $a^*_x \in \mathcal{A}_{\hat{\pi}_x}$. Assume for contradiction that $O_{\mathcal{M}_y}(\phi(s^*_x), \psi(a^*_x)) = 1$. Then, $\psi(a^*_x) \in \text{supp}(\pi_y(\cdot|\phi(s^*_x))$ since π_y is a covering optimal policy from Definition 5, which implies $g^{-1}(a^*_x) \in \text{supp}(\pi_y(\cdot|f(s^*_x))$ since $\sigma^x_{\hat{\pi}_x}(s^*_x) > 0$ and $a^*_x \in \mathcal{A}_{\hat{\pi}_x}$. It follows that $g(g^{-1}(a^*_x)) \in \text{supp}(g(\pi_y(\cdot|f(s^*_x)))) \Rightarrow a^*_x \in \text{supp}(\hat{\pi}_x(\cdot|s^*_x))$ since $\hat{\pi}_x(\cdot|s^*_x)$ is the pushforward measure $g(\pi_y(\cdot|f(s^*_x)))$. Then, $\sigma^x_{\hat{\pi}_x}(s^*_x, a^*_x) \overset{\text{Eq25}}{=} \sigma^x_{\hat{\pi}_x}(s^*_x)\hat{\pi}_x(a^*_x|s^*_x) > 0$, since $\hat{\pi}_x(a^*_x|s^*_x) > 0$ and $\sigma^x_{\hat{\pi}_x}(s^*_x) > 0$, which contradicts $(s^*_x, a^*_x) \notin \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x))$. This concludes the proof of Equation 28.

We proceed to show that the optimal policy and dynamics preservation properties hold for our construction (ϕ, ψ).

• Optimality (Eq. 1): From the converse direction of the above subclaim and the optimality of $\hat{\pi}_x$ the result immediate follows:

$$\begin{aligned}
O_{\mathcal{M}_y}(\phi(s^*_x), \psi(a^*_x)) = 1 &\overset{\text{Eq28}}{\Rightarrow} (s^*_x, a^*_x) \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x)) \\
&\overset{\text{Eq25}}{\Rightarrow} \hat{\pi}_x(a^*_x|s^*_x) > 0 \\
&\overset{\text{Cor1}}{\Rightarrow} O_{\mathcal{M}_x}(s^*_x, a^*_x) = 1 \ \ \forall s^*_x \in \mathcal{S}_x, a^*_x \in \mathcal{A}_x
\end{aligned}$$

• Surjection (Eq. 2): Assume for contradiction $\exists(s_y^*, a_y^*)$ such that $O_{\mathcal{M}_y}(s_y^*, a_y^*) = 1$, but $\phi^{-1}(s_y^*) = \emptyset$ or $\psi^{-1}(a_y^*) = \emptyset$. Since $O_{\mathcal{M}_y}(s_y^*, a_y^*) = 1$ we have $s_y^* \neq s_y^d, a_y^* \neq a_y^d$. Thus $\phi(s_y^*)^{-1} = f^{-1}(s_y^*)$ and $\psi^{-1}(a_y^*) = \{(g^{-1})^{-1}(a_y^*)\} = \{g(a_y^*)\}$. Since g is a function defined $\forall a_y \in \mathcal{A}_y$, it follows that $\psi^{-1}(a_y^*) \neq \emptyset$. Thus it must be that $\phi^{-1}(s_y^*) = \emptyset$. Let $s_y^{*'} = P_y(s_y^*, a_y^*)$. Then,

$$
\begin{aligned}
\sigma^{x\to y}_{\hat{\pi}_x}(s_y^*, a_y^*, s_y^{*'}) &\overset{\text{Lemma 4}}{=} \lim_{t\to\infty} \Pr(\hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*, \hat{s}_y^{(t+1)} = s_y^{*'}) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t+1)} = s_y^{*'} | \hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*) \Pr(\hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t+1)} = s_y^{*'} | \hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*) \Pr(\hat{a}_y^{(t)} = a_y^* | \hat{s}_y^{(t)} = s_y^*) \Pr(\hat{s}_y^{(t)} = s_y^*) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t+1)} = s_y^{*'} | \hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*) \pi_y(a_y^* | s_y^*) \textstyle\sum_{s_x \in \phi^{-1}(s_y^*)} \Pr(\hat{s}_x^{(t)} = s_x) \\
&= \lim_{t\to\infty} \Pr(\hat{s}_y^{(t+1)} = s_y^{*'} | \hat{s}_y^{(t)} = s_y^*, \hat{a}_y^{(t)} = a_y^*) \pi_y(a_y^* | s_y^*) \cdot 0 \\
&= 0
\end{aligned}
$$

However,

$$
\begin{aligned}
\sigma^{y}_{\pi_y}(s_y^*, a_y^*, s_y^{*'}) &\overset{\text{Eq 27}}{=} \mathbb{1}(P_y(s_y^*, a_y^*) = P_y(s_y^*, a_y^*)) \pi_y(a_y^* | s_y^*) \sigma^y_{\pi_y}(s_y^*) \\
&= \pi_y(a_y^* | s_y^*) \sigma^y_{\pi_y}(s_y^*) > 0
\end{aligned}
$$

To see why the last inequality holds, first recall that $\mathcal{M}_y$ is unichain and π_y is stochastic optimal for $\mathcal{M}_y$, so the stationary distribution over states have full support over $\mathcal{S}_y$ ($\because$ stationary distributions of irreducible markov chains are fully supported over the entire state space) Therefore $\sigma^y_{\pi_y}(s_y) \overset{\text{Lemma4}}{=} \lim_{t\to\infty} \Pr(s_y^{(t)} = s_y; \pi_y, P_y) > 0 \quad \forall s_y \in \mathcal{S}_y$. Thus, we have $\sigma^y_{\pi_y}(s_y^*) > 0$. Furthermore, $\pi_y(a_y^* | s_y^*) > 0$ by Corollary 1. Putting these two results together, we obtain $\sigma^y_{\pi_y}(s_y^*, a_y^*, s_y^{*'}) > 0$. Then, $\sigma^{x\to y}_{\hat{\pi}_x} \neq \sigma^y_{\pi_y}$ which contradicts the satisfiability of objective 3.

• Dynamics (Eq. 3): Assume for contradiction that $\exists s_x^-, a_x^-$ and $s_x^{-'} = P_x(s_x^-, a_x^-$ such that $O_{\mathcal{M}_y}(\phi(s_x^-), \psi(a_x^-)) = 1$ but the dynamics preservation property is violated, i.e $P_y(\phi(s_x^-), \psi(a_x^-)) \neq \phi(P_x(s_x^-, a_x^-)) = \phi(s_x^{-'})$. If $(s_x^-, a_x^-) \notin \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x))$, then $O_{\mathcal{M}_y}(\phi(s_x^-), \psi(a_x^-)) = 0$ by Equation 28 which contradicts $O(\phi(s_x^-), \psi(a_x^-)) = 1$. Thus, it must be that $(s_x^-, a_x^-) \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x))$ which further implies $(s_x^-, a_x^-, s_x^{-'}) \in \text{supp}(\sigma^x_{\hat{\pi}_x}(s_x, a_x, s'_x))$ by Equation 26 and $\phi(s_x^-) = f(s_x^-), \psi(a_x^-) = g^{-1}(a_x^-)$ by Equation 25 since $\sigma^x_{\hat{\pi}_x}(s_x^-) > 0, \hat{\pi}(a_x^- | s_x^-) > 0$.

Let $\mathcal{F} : \mathcal{S}_x \times g(\mathcal{A}_y) \times \mathcal{S}_x \to \mathcal{S}_y \times \mathcal{A}_y \times \mathcal{S}_y$ be a function $(a, b, c) \mapsto (f(a), g^{-1}(b), f(c))$. Then, by Lemma 6, we have $\sigma^{x\to y}_{\hat{\pi}_x}(s_x, a_x, s'_x) = \mathcal{F}(\rho^x_{\hat{\pi}_x}(s_x, a_x, s'_x)) \overset{\text{Lemma4}}{=} \mathcal{F}(\sigma^x_{\hat{\pi}_x}(s_x, a_x, s'_x))$. So,

$$
\sigma^x_{\hat{\pi}_x}(s_x^-, a_x^-, s_x^{-'}) > 0 \Rightarrow \sigma^{x\to y}_{\hat{\pi}_x}(\mathcal{F}(s_x^-, a_x^-, s_x^{-'})) = \sigma^{x\to y}_{\hat{\pi}_x}(f(s_x^-), g^{-1}(a_x^-), f(s_x^{-'})) > 0
$$

Thus, $(f(s_x^-), g^{-1}(a_x^-), f(s_x^{-'})) = (\phi(s_x^-), \psi(a_x^-), \phi(s_x^{-'})) \in \text{supp}(\sigma^{x\to y}_{\hat{\pi}_x}(s_x, a_x, s'_x))$. However,

$$
\begin{aligned}
\sigma^y_{\pi_y}(\phi(s_x^-), \psi(a_x^-), \phi(s_x^{-'})) &\overset{\text{Eq 26}}{=} \sigma^y_{\pi_y}(\phi(s_x^-), \psi(a_x^-)) \mathbb{1}\big(\phi(s_x^{-'}) = P_y(\phi(s_x^-), \psi(a_x^-))\big) \\
&= \sigma^y_{\pi_y}(\phi(s_x^-), \psi(a_x^-)) \cdot 0 \\
&= 0
\end{aligned}
$$

Thus, $\text{supp}(\sigma^{x\to y}_{\hat{\pi}_x}) \neq \text{supp}(\sigma^y_{\pi_y}) \Rightarrow \sigma^{x\to y}_{\hat{\pi}_x} \neq \sigma^y_{\pi_y}$ which contradicts f, g satisfying objective 3. This concludes the proof of the main theorem. □