Appendices

A Introduction to do-calculus

The framework of *do*-calculus [30] was proposed as an intuitive tool to answer identifiability questions given a causal graph $\mathcal{G}$, such as, can the interventional distribution $p(y|do(x), z)$ be recovered from the observational distributions $p(y, x, z)$?

A.1 The three rules of do-calculus

Do-calculus relies on three graphical rules, which depend solely on the existence of specific structural constraints in G:

- R1: insertion/deletion of observations, $p(y|do(x), z, w) = p(y|do(x), w)$ if Y and Z are d-separated by $X \cup W$ in $\mathcal{G}^\star$, the graph obtained from $\mathcal{G}$ by removing all arrows pointing into variables in X.
- R2: action/observation exchange, $p(y|do(x), do(z), w) = p(y|do(x), z, w)$ if Y and Z are d-separated by $X \cup W$ in $\mathcal{G}^\dagger$, the graph obtained from $\mathcal{G}$ by removing all arrows pointing into variables in X and all arrows pointing out of variables in Z.
- R3: insertion/deletion of actions, $p(y|do(x), do(z), w) = p(y|do(x), w)$ if Y and Z are d-separated by $X \cup W$ in $\mathcal{G}^\ddagger$, the graph obtained from $\mathcal{G}$ by first removing all the arrows pointing into variables in X (thus creating $\mathcal{G}^\star$) and then removing all of the arrows pointing into variables in Z that are not ancestors of any variable in W in $\mathcal{G}^\star$.

This set of rules has been shown to be complete [13; 33], and results in an algorithm polynomial in the number of nodes in $\mathcal{G}$ to answer identifiability questions, which either outputs "no" or "yes" along with an estimate (a recovery formula) based on observational quantities. We refer the reader to Pearl [30] for a thorough introduction to *do*-calculus.

A.2 Note on ignorability and exogeneity

In this paper we use at great length the concept of confounding, which is a core idea in Judea Pearl's *do*-calculus framework. For readers who are more familiar with the framework of potential outcomes from Donald Rubin [14], the concept of confounding closely relates to the concepts of ignorability and exogeneity, which can be shown to be equivalent to the unconfoundedness (no confounding) assumption [28].

B Experimental details

B.1 Training

In all our experiments we use tabular logistic models for each of the components in $\hat{q}$. That is, each building bloc $q(z_0)$, $q(o_t|z_t)$, $q(z_{t+1}|z_t, a_t)$, and $q(a_t|h_t, z_t, i = 0)$ is parameterized using a set of softmax-normalized scalars vectors. We train $\hat{q}$ via gradient descent using the Adam optimizer [17], by directly minimizing the negative log likelihood of the model (equation (5)) on random mini-batches of trajectories sampled from $\mathcal{D}_{std} \cup \mathcal{D}_{prv}$. Agents are trained using the learned model as a "dream" environment (by sampling imaginary trajectories $\tau \sim \hat{q}(\tau|i = 1)$), with a simple actor-critic algorithm (REINFORCE with a state-value baseline) for a fixed number of iterations, also using the Adam optimizer. Both the actor and critic consists of a 2-layers perceptron (MLP) with the same hidden layer size, which take as input the belief state recovered from the model. The training hyperparameters we used in each experiment are displayed in table 1.

	`tiger`	`hidden treasures`	`sloppy dark room`
Latent model			
latent space size $\lvert\mathcal{Z}\rvert$	32	256	128
learning rate	10^{-3}	10^{-3}	10^{-3}
number of epochs (max)	500	500	500
number of gradient steps per epoch	50	100	100
minibatch size (trajectories τ)	32	64	64
Actor-critic agent			
exploration noise ϵ	0.5	0.2	0.2
hidden layer size	256	512	256
learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}
number of epochs	200	400	200
number of gradient steps per epoch	50	50	50
minibatch size (trajectories τ)	32	64	64
minibatch return scaling	yes	no	no
entropy bonus	10^{-3}	10^{-3}	10^{-3}
discount factor γ	1	1	1

Table 1: Training hyperparameters we used in each experiment. When learning the model, we divide the learning rate by 10 after 10 epochs without loss improvement (reduce on plateau), and we stop training after 20 epochs without improvement (early stopping). We use all available data for training, and we monitor the training loss for early stopping (no validation set).

B.2 Evaluation

Model quality (likelihood). To evaluate the general quality of the recovered POMDP model, we compute the likelihood of $\hat{q}$ on a new interventional dataset $\mathcal{D}_{test}$ obtained from the true environment p with a uniformly random policy π_{rand},

$$\mathbb{E}_{\tau \sim p_{init}, p_{trans}, p_{obs}, \pi_{rand}} \left[\hat{q}(o_0) \prod_{t=1}^{|\tau|} \hat{q}(o_{t+1}|h_t, i = 1) \right].$$

We report an empirical estimate of this measure using 10000 trajectories.

Agent performance (cumulated reward). To evaluate quality of the agent obtained from the model $\hat{q}$ for solving the standard POMDP control task, we compute the expected cumulated reward of the policy $\hat{\pi}^\star$

on the true environment p,

$$\mathbb{E}_{\tau \sim p_{init}, p_{trans}, p_{prv}, \hat{\pi}^\star}\left[\sum_{t=1}^{|\tau|} r(o_t)\right].$$

We report an empirical estimate of this measure using 10000 trajectories.

B.3 Tiger experiment

We present the (compact) POMDP dynamics of the `tiger` problem in table 2. After conversion to the notation in the paper, the observations become $o_t = (roar_t, reward_t)$, the actions remain $a_t = action_t$, and the hidden states are $s_t = (tiger_t, reward_t)$. The privileged policies used in the experiments (section 5.2) are reported in table 3.

Table 2: Compact POMDP dynamics in the `tiger` problem.

$tiger_0$	
left	right
0.5	0.5

$p(tiger_0)$

	$roar_t$	
$tiger_t$	left	right
left	0.85	0.15
right	0.15	0.85

$p(roar_t \mid tiger_t)$

		$tiger_{t+1}$	
$tiger_t$	$action_t$	left	right
left	listen	1.0	0.0
	open left	0.5	0.5
	open right	0.5	0.5
right	listen	0.0	1.0
	open left	0.5	0.5
	open right	0.5	0.5

$p(tiger_{t+1} \mid tiger_t, action_t)$

		$reward_{t+1}$		
$tiger_t$	$action_t$	-1	-100	+10
left	listen	1.0	0.0	0.0
	open left	0.0	1.0	0.0
	open right	0.0	0.0	1.0
right	listen	1.0	0.0	0.0
	open left	0.0	0.0	1.0
	open right	0.0	1.0	0.0

$p(reward_{t+1} \mid tiger_t, action_t)$

Table 3: Privileged policies $\pi_{prv}(action \mid tiger)$ used in the `tiger` experiment.

		$action_t$		
privileged policy	$tiger_t$	listen	left	right
random	left	0.33	0.33	0.33
	right	0.33	0.33	0.33
noisy good	left	0.05	0.30	0.65
	right	0.05	0.80	0.15
perfect good	left	0.00	0.00	1.00
	right	0.00	1.00	0.00
perfect bad	left	0.00	1.00	0.00
	right	0.00	0.00	1.00

C Additional empirical results

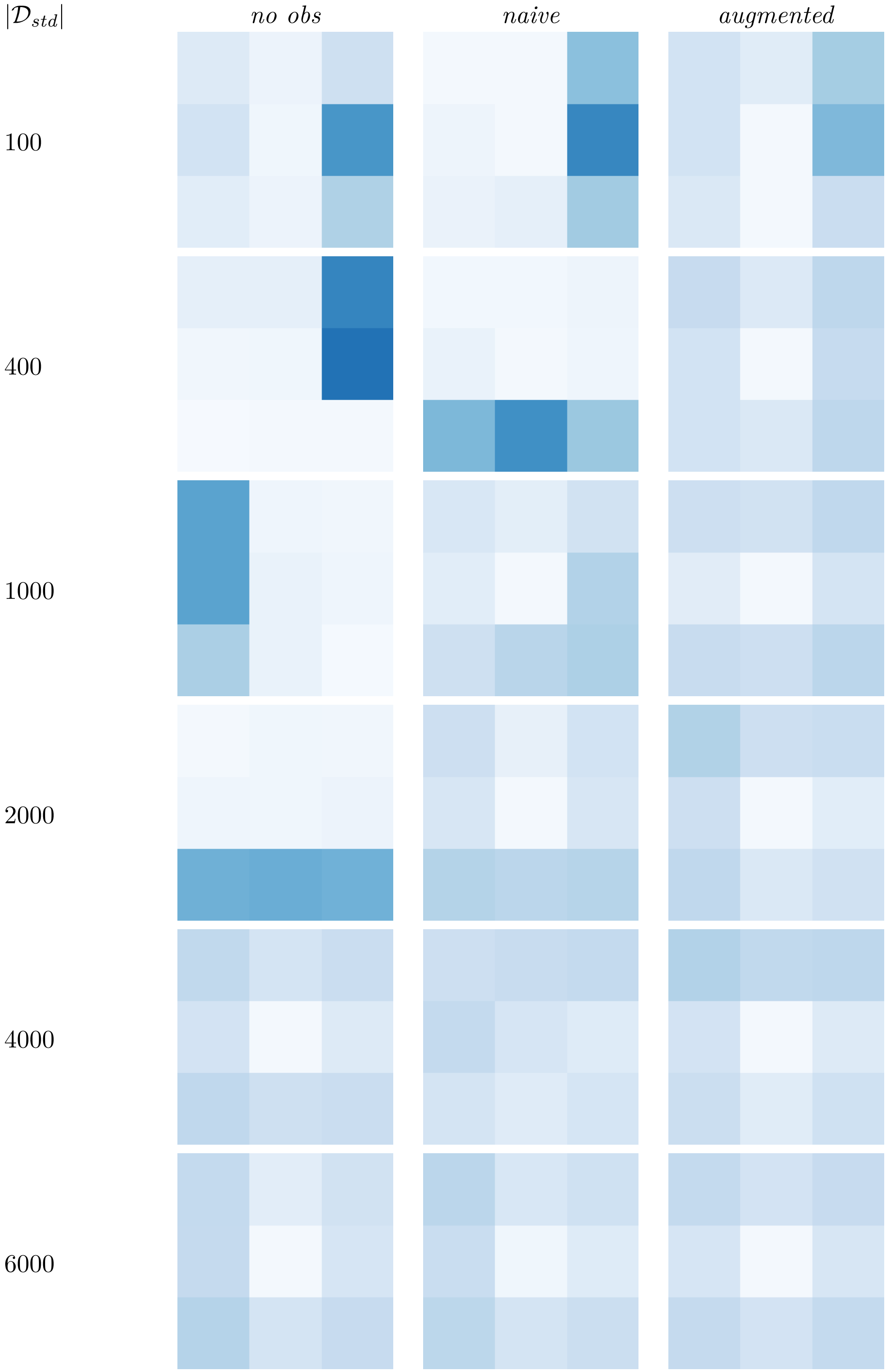

Figure 8: Evolution of the test-time agent trajectories in the `hidden treasures` experiment. We report a heatmap of the tiles visited by each agent (*no obs*, *naive*, *augmented*) at different time steps (number of interventional samples collected) during a single RL run (single seed). Eventually all methods converge to the optimal strategy, which is to cycle through the 4 corners. Our *augmented* method converges to this behaviour earlier on during training.

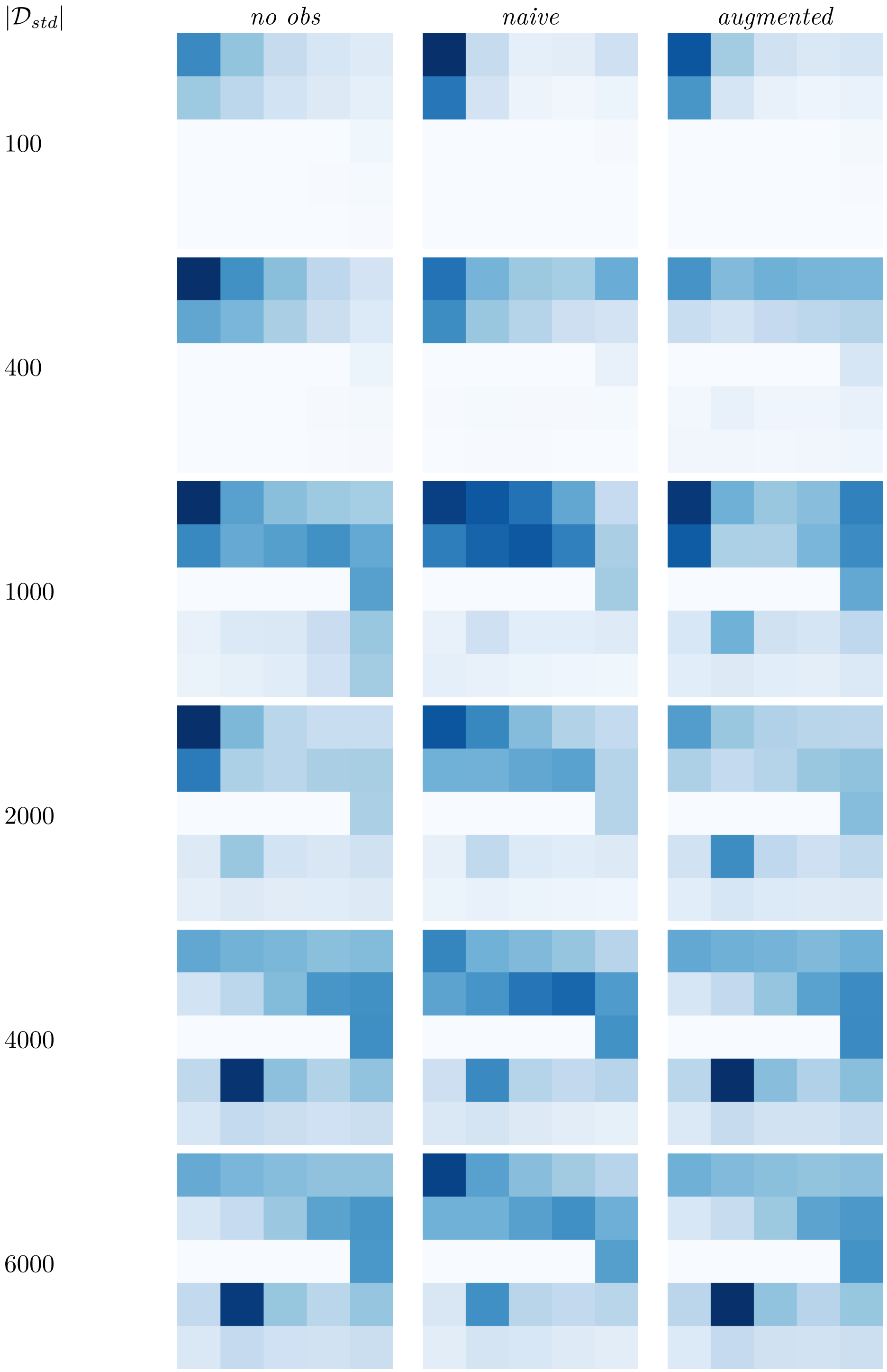

Figure 9: Evolution of the test-time agent trajectories in the `sloppy dark room` experiment. We report a heatmap of the tiles visited by each agent (*no obs*, *naive*, *augmented*) at different time steps (number of interventional samples collected), averaged over 10 RL runs (10 seeds). Eventually all methods manage to consistently overcome the obstacle and reach the target tile. Our *augmented* method converges to this behaviour earlier on during training.

D Proof of Theorem 1.

Theorem 1. *Assuming $|\mathcal{D}_{prv}| \to \infty$, for any $\mathcal{D}_{std}$ the recovered causal model is bounded as follows:*

$$\prod_{t=0}^{T-1} \hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})) \geq \prod_{t=0}^{T-1} p(a_t|h_t, i=0)p(o_{t+1}|h_t, a_t, i=0), \text{ and}$$
$$\prod_{t=0}^{T-1} \hat{q}(o_{t+1}|o_{0\to t}, do(a_{0\to t})) \leq \prod_{t=0}^{T-1} p(a_t|h_t, i=0)p(o_{t+1}|h_t, a_t, i=0) + 1 - \prod_{t=0}^{T-1} p(a_t|h_t, i=0),$$

$\forall h_{T-1}, a_{T-1}, T \geq 1$ *where* $p(h_{T-1}, a_{T-1}, i=0) > 0$.

Proof of Theorem 1. Consider $q(\tau, i) \in \mathcal{Q}$ any distribution that follows our augmented POMDP constraints. As an intermediary step, we will start by proving the following

$$\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i=1) = \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i=0) \prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, a_t, h_t, i=0). \tag{6}$$

First, for any $0 \leq t \leq T-1$, we can write the following factorization

$$q(z_t, z_{t+1}, o_{t+1}|h_t, a_t, i=1) = q(z_t|h_t, a_t, i=1)q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=1).$$

Because of the augmented POMDP constraints, the independences $Z_t \perp\!\!\!\perp A_t \mid H_t, I=1$ and $Z_{t+1}, O_{t+1} \perp\!\!\!\perp I \mid Z_t, A_t, H_t$ hold in q, which further allows us to write

$$q(z_t, z_{t+1}, o_{t+1}|h_t, a_t, i=1) = q(z_t|h_t, i=1)q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=0). \tag{7}$$

Then, we directly get

$$q(o_{t+1}|h_t, a_t, i=1) = \sum_{z_t, z_{t+1}}^{\mathcal{Z}\times\mathcal{Z}} q(z_t|h_t, i=1)q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=0). \tag{8}$$

Now, let us consider the special case where $T=1$. We can use the constraint $Z_0 \perp\!\!\!\perp I \mid H_0$ to write

$$q(o_1|h_0, a_0, i=1) = \sum_{z_{0\to 1}}^{\mathcal{Z}^2} q(z_0|h_0, i=0)q(z_1, o_1|z_0, h_0, a_0, i=0),$$

which is equation (6), the desired result, for $T=1$. In the case where $T \geq 2$, we can reuse equation (8) to write

$$\begin{aligned}
q(o_T|h_{T-1}, a_{T-1}, i=1) &= \sum_{z_{T-1\to T}}^{\mathcal{Z}^2} q(z_{T-1}|h_{T-2}, a_{T-2}, o_{T-1}, i=1)q(z_T, o_T|z_{T-1}, h_{T-1}, a_{T-1}, i=0) \\
&= \sum_{z_{T-1\to T}}^{\mathcal{Z}^2} \frac{q(z_{T-1}, o_{T-1}|h_{T-2}, a_{T-2}, i=1)}{q(o_{T-1}|h_{T-2}, a_{T-2}, i=1)} q(z_T, o_T|z_{T-1}, h_{T-1}, a_{T-1}, i=0) \\
\prod_{t=T-2}^{T-1} q(o_{t+1}|h_t, a_t, i=1) &= \sum_{z_{T-1\to T}}^{\mathcal{Z}^2} q(z_{T-1}, o_{T-1}|h_{T-2}, a_{T-2}, i=1)q(z_T, o_T|z_{T-1}, h_{T-1}, a_{T-1}, i=0).
\end{aligned}$$

Then, we can introduce variable Z_{T-2} and use equation (7) again to obtain

$$\begin{aligned}
\prod_{t=T-2}^{T-1} q(o_{t+1}|h_t, a_t, i=1) &= \sum_{z_{T-2\to T}}^{\mathcal{Z}^3} q(z_{T-2}, z_{T-1}, o_{T-1}|h_{T-2}, a_{T-2}, i=1)q(z_T, o_T|z_{T-1}, h_{T-1}, a_{T-1}, i=0) \\
&= \sum_{z_{T-2\to T}}^{\mathcal{Z}^3} q(z_{T-2}|h_{T-2}, i=1) \prod_{t=T-2}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i=0).
\end{aligned}$$

In the case where $T = 2$, we can use $Z_0 \perp\!\!\!\perp I \mid H_0$ again to obtain equation (6), the desired result for $T = 2$. In the case where $T \geq 3$, we can apply the same steps again to obtain

$$\prod_{t=T-3}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) = \sum_{z_{T-3\to T}}^{\mathcal{Z}^4} q(z_{T-3}|h_{T-3}, i = 1) \prod_{t=T-3}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0).$$

Now, either $T = 3$ and we can use $Z_0 \perp\!\!\!\perp I \mid H_0$ to obtain equation (6), or $T \geq 4$ and we can continue the decomposition by introducing Z_{T-4}. By following this recursive approach we eventually reach Z_0 and prove equation (6) for any T.

Let us now re-express equation (6) as follows

$$\begin{aligned}\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) = &\sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \left(\prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0)\right) \left(\prod_{t=0}^{T-1} q(a_t|z_t, h_t, i = 0)\right) \\ &+ \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \left(\prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0)\right) \left(1 - \prod_{t=0}^{T-1} q(a_t|z_t, h_t, i = 0)\right)\Bigg) \\ \prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) = &\prod_{t=0}^{T-1} q(a_t|h_t, i = 0) q(o_{t+1}|h_t, a_t, i = 0) \\ &+ \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \left(\prod_{t=0}^{T-1} q(z_{t+1}, o_{t+1}|z_t, h_t, a_t, i = 0)\right) \left(1 - \prod_{t=0}^{T-1} q(a_t|z_t, h_t, i = 0)\right).\end{aligned}$$

By assuming probabilities are positive, we can substitute the second term by 0 to obtain our lower bound

$$\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) \geq \prod_{t=0}^{T-1} q(a_t|h_t, i = 0) q(o_{t+1}|h_t, a_t, i = 0).$$

Then by assuming probabilities are upper bounded by 1, we can substitute $q(o_{t+1}|z_{t+1}, z_t, h_t, a_t, i = 0)$ by 1 to obtain our upper bound

$$\begin{aligned}\prod_{t=0}^{T-1} q(o_{t+1}|h_t, a_t, i = 1) \leq &\prod_{t=0}^{T-1} q(a_t|h_t, i = 0) q(o_{t+1}|h_t, a_t, i = 0) \\ &+ \sum_{z_{0\to T}}^{\mathcal{Z}^{T+1}} q(z_0|h_0, i = 0) \left(\prod_{t=0}^{T-1} q(z_{t+1}|z_t, h_t, a_t, i = 0)\right) \left(1 - \prod_{t=0}^{T-1} q(a_t|z_t, h_t, i = 0)\right) \\ \leq &\prod_{t=0}^{T-1} q(a_t|h_t, i = 0) q(o_{t+1}|h_t, a_t, i = 0) + 1 - \prod_{t=0}^{T-1} q(a_t|h_t, i = 0).\end{aligned}$$

Finally, with $\hat{q}$ solution of (5) and $|\mathcal{D}_{prv}| \to \infty$ we have that $D_{\mathrm{KL}}(p(\tau|i = 0)\|\hat{q}(\tau|i = 0)) = 0$, and thus $\hat{q}(a_t|h_t, i = 0) = p(a_t|h_t, i = 0)$ and in particular $\hat{q}(o_{t+1}|h_t, a_t, i = 0) = p(o_{t+1}|h_t, a_t, i = 0)$, which allows us to conclude. □