Reinformed Dreamer: An Asymmetric World Model Efficiently Trained through Latent Guidance

Gaspard Lambrechts[1,2], **Adrien Bolland**[3], **Daniel Ebi**[4], **Damien Ernst**[3]
gaspard.lambrechts@mcgill.ca

[1]**McGill University, Canada**
[2]**Mila Québec Artificial Intelligence Institute, Canada**
[3]**Université de Liège, Belgium**
[4]**Karlsruhe Institute of Technology, Germany**

Abstract

Much like humans benefit from guidance while learning, reinforcement learning algorithms may benefit from additional supervision beyond rewards. Leveraging additional information during training to learn better representations and behaviors has been the focus of asymmetric reinforcement learning. This learning paradigm has proven effective under partial observability when additional state information is available, but also under full observability when more refined state information is available. Focusing on model-based reinforcement learning, we study the effect of asymmetric learning on observation representations and on privileged information representations. First, we identify a limitation in the privileged information representations learned by an asymmetric model-based algorithm known as the Informed Dreamer. Then, we propose a novel asymmetric representation learning objective using latent guidance, resulting in a new algorithm called the Reinformed Dreamer. Experiments across several benchmarks show a more consistent improvement over Dreamer than previous asymmetric approaches.

1 Introduction

In the field of artificial intelligence, intelligence is usually understood as the ability to make decisions, based on perception, in order to achieve goals (McCarthy, 1998; Russell, 2010). In other words, intelligence is about perceiving and abstracting past information about the world for then acting on its future execution, in the perspective of achieving objectives (Lambrechts, 2025). Reinforcement learning (RL) is an appealing approach for learning intelligent behaviors, notably because it makes few assumptions about the decision problem (Sutton & Barto, 1998). In its purest form, the promise of an RL algorithm is to learn an optimal behavior from interaction with an unknown decision process. In general, the perception of the decision process consists of partial observations of its underlying state (Åström, 1965), which defines a partially observable Markov decision process (POMDP). Under partial observability, the optimal decision generally depends on the entire history of observations and past actions. These histories grow unboundedly and demand an appropriate method to compress them into adequate representations. Learning these abstractions, or representations, is known as representation learning, and has been one of the main focuses of RL in POMDP (Ni et al., 2024).

While learning from interaction and solely from partial observations of the decision process is a great promise, it is a restrictive learning paradigm. As initially noticed in the supervised learning setting (Vapnik & Vashist, 2009), additional information may be available during training to improve learning. In RL, similar ideas have emerged under the name asymmetric RL (Pinto et al., 2018), motivated by the asymmetry of observability between training and execution. The additional information may take the form of privileged observations from additional sensors, access to estimates of the state or to the simulator state, information about future states in hindsight, etc. Recently, a line of work has focused on proposing theoretically grounded asymmetric learning objectives (Warrington et al., 2021; Baisero & Amato, 2022; Lambrechts et al., 2024; Wang et al., 2023). Some of these methods, especially those leveraging an asymmetric critic, have demonstrated impressive empirical performance in various domains (Degrave et al., 2022; Kaufmann et al., 2023; Vasco et al., 2024; Hu et al., 2025; Dürr et al., 2026). We provide an overview of asymmetric RL methods in Appendix A.

In the symmetric setting, model-based RL has proven an effective approach for solving decision-making problems, including under partial observability (Hafner et al., 2025; Samsami et al., 2024). In model-based RL, a model of the decision process, known as a world model (Ha & Schmidhuber, 2018), is learned from interaction samples. In the Dreamer algorithm (Hafner et al., 2020; 2021; 2025), which we will refer to as the Uninformed Dreamer in the following, the world model is trained to recurrently predict a latent representation of the reward and next observation given the history of observations and actions. Such world modeling objectives correspond to representation learning objectives that provide representations of the history that are sufficient for optimal control (Striebel, 1965), also known as information states (Subramanian et al., 2022). From the world model, and using an adequate policy, it is possible to generate synthetic trajectories of interaction using these latent representations, a process called imagination. It has proved effective to use an actor-critic algorithm based on these synthetic trajectories in order to optimize the policy (Hafner et al., 2025).

Recently, model-based algorithms were adapted to the asymmetric setting to benefit from eventual additional information. The Informed Dreamer algorithm (Lambrechts et al., 2024) adapted the Uninformed Dreamer (Hafner et al., 2025) by introducing an asymmetric world model trained to predict a latent representation of the reward and next privileged information given the history of observations and actions, instead of a representation of the reward and next observation. By ensuring that the privileged information embeds all relevant information from the observation, its representation is also a sufficient representation of the observation, and the imagination process remains valid. This algorithm demonstrated an improved learning speed on various partially observable environments. The Scaffolder algorithm (Hu et al., 2024) outperformed the Informed Dreamer by learning a privileged world model in addition. The privileged world model learns a latent representation of the reward and next privileged information based on the history of privileged informations and past actions. To obtain the representations based on the history of observations and past actions, needed during execution, an unprivileged world model is learned jointly. During imagination, the privileged world model first generates latent representations based on the privileged history. These representations are then decoded into observations and reencoded by the unprivileged world model. This results in a costly imagination process, and one would ideally aim to achieve similar performance with a single world model and without explicitly decoding the observations.

This work makes several contributions towards this goal. First, we show that the learning objective for the latent representation of the privileged information is problematic in the Informed Dreamer. Indeed, the asymmetric variational autoencoder uses a generally loose lower bound on the likelihood of the privileged information. Second, we propose a proper learning objective for the latent representation of the privileged information by introducing a privileged variational encoder. Since the privileged encoder cannot be used during execution, an unprivileged encoder is jointly learned through latent guidance. This new world modeling objective defines the Reinformed Dreamer algorithm, which features two encoders in a single world model, and whose imagination process remains unchanged. Third, we show in several suites of environments that the Reinformed Dreamer outperforms or matches previous approaches in terms of convergence speed and final performance.

2 Informed Partially Observable Markov Decision Processes

We model the asymmetry of observability between training and execution with the informed POMDP (Lambrechts et al., 2024). It is a generalization of the POMDP (Åström, 1965) that introduces the (privileged) information that is available during training. Formally, an informed POMDP $\widetilde{\mathcal{P}}$ is a tuple $\widetilde{\mathcal{P}} = (\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}, P, T, R, \widetilde{I}, \widetilde{O}, \gamma)$, where $\mathcal{S}, \mathcal{A}, \mathcal{I}, \mathcal{O}$ are the state, action, information, and observation spaces, respectively. $P \in \Delta(\mathcal{S})$ is the initial state distribution that gives the probability $P(s_0)$ of $s_0 \in \mathcal{S}$ being the initial state. $T\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition distribution that gives the probability $T(s_{t+1}|s_t, a_t)$ of $s_{t+1} \in \mathcal{S}$ being the state resulting from action $a_t \in \mathcal{A}$ in state $s_t \in \mathcal{S}$. $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function that gives the expected immediate reward $r_t = R(s_t, a_t)$ obtained when taking action $a_t \in \mathcal{A}$ in state $s_t \in \mathcal{S}$. $\widetilde{I}$ is the information distribution that gives the probability $\widetilde{I}(i_t|s_t)$ of obtaining information $i_t \in \mathcal{I}$ in state $s_t \in \mathcal{S}$, and $\widetilde{O}$ is the observation distribution that gives the probability $\widetilde{O}(o_t|i_t)$ of obtaining observation $o_t \in \mathcal{O}$ given information $i_t \in \mathcal{I}$. $\gamma \in [0, 1)$ is the discount factor that weighs the relative importance of future rewards.

Taking a sequence of t actions in the informed POMDP conditions its execution and provides samples $(i_0, o_0, a_0, r_0, \ldots, i_t, o_t)$ during training. During execution, however, the privileged informations and rewards are not available anymore. As a result, interacting with the informed POMDP becomes equivalent to interacting with its underlying execution POMDP, defined as $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, T, R, O, \gamma)$ where $O(o_t|s_t) = \mathbb{E}_{\widetilde{I}(i|s_t)}[\widetilde{O}(o_t|i)]$ is the observation distribution. Taking a sequence of t actions in the execution POMDP conditions its execution and provides the history $h_t = (o_0, a_0, \ldots, o_t) \in \mathcal{H}$, where $\mathcal{H}$ is the set of histories of arbitrary length. Note that the information samples $i_0, \ldots, i_t$ and reward samples $r_0, \ldots, r_{t-1}$ are not included. The training and execution are summarized in Figure 1.

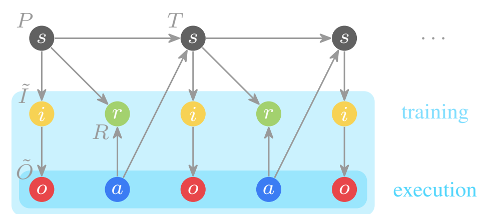


Figure 1: Bayesian graph of an informed POMDP (Lambrechts et al., 2024).

A history-dependent stochastic policy $\eta \in \mathrm{H}$ is a mapping from histories to probability distributions over the action space, where $\mathrm{H} = \{\eta\colon \mathcal{H} \to \Delta(\mathcal{A})\}$ is the set of such mappings. A policy η^* is said to be optimal for an informed POMDP when it is optimal in the underlying execution POMDP, i.e., when it maximizes the expected return $J(\eta) = \mathbb{E}^{\eta}[\sum_{t=0}^{\infty} \gamma^t r_t]$. The RL objective for an informed POMDP is thus to find an optimal policy $\eta^* \in \arg\max_{\eta \in \mathrm{H}} J(\eta)$ from online or offline interaction, that is from samples $\left\{(i_0^n, o_0^n, a_0^n, r_0^n, \ldots, i_{t_n}^n, o_{t_n}^n)\right\}_{n=0}^{N-1}$.

3 Informed Dreamer and Information Gap

In Subsection 3.1, we recall the Informed Dreamer (Lambrechts et al., 2024), an asymmetric model-based RL algorithm for informed POMDPs, and the Uninformed Dreamer (Hafner et al., 2025) is obtained as a special case of this algorithm when there is no additional information (i.e., when $i_t = o_t$). In Subsection 3.2, we show that the learning objective of the world model of the Informed Dreamer is optimized through a generally loose lower bound.

3.1 Informed Dreamer

World Model. In an informed POMDP, an informed world model is a model $p_\phi(r_t, i_{t+1}|h_t, a_t)$ of the history-dependent distribution $p(r_t, i_{t+1}|h_t, a_t)$. In the Informed Dreamer, the distribution $p_\phi(r_t, i_{t+1}|h_t, a_t)$ is implemented by a particular latent variable model (LVM), known as a recurrent state space model (RSSM). More precisely, for an initial hidden state $z_0 = 0$, at any $t \geq 0$,

$$e_k \sim q^e_\phi(e_k|z_k, o_k),\ k \leq t, \tag{encoder, 1}$$
$$z_{k+1} = u_\phi(z_k, e_k, a_k),\ k \leq t, \tag{recurrence, 2}$$
$$\hat{e}_{t+1} \sim p^{\hat{e}}_\phi(\hat{e}_{t+1}|z_{t+1}), \tag{prior, 3}$$
$$\hat{r}_t \sim p^r_\phi(\hat{r}_t|z_{t+1}, \hat{e}_{t+1}), \tag{reward decoder, 4}$$
$$\hat{i}_{t+1} \sim p^i_\phi(\hat{i}_{t+1}|z_{t+1}, \hat{e}_{t+1}). \tag{information decoder, 5}$$

This LVM uses latent variables $(e_{0:t}, \hat{e}_{t+1})$ and its prior $p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)$ is given by:

$$q_\phi(e_{0:t}|h_t) = \textstyle\prod_{k=0}^{t} q^e_\phi(e_k|z_k, o_k), \tag{6}$$

$$p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t) = q_\phi(e_{0:t}|h_t) p^{\hat{e}}_\phi(\hat{e}_{t+1}|z_{t+1}), \tag{7}$$

where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$. The distribution $p_\phi(r_t, i_{t+1}|h_t, a_t)$ is thus modeled by:

$$\begin{aligned} p_\phi(r_t, i_{t+1}|h_t, a_t) &= \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} [p_\phi(r_t, i_{t+1}|z_{t+1}, \hat{e}_{t+1})] && (8) \\ &= \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} [p^r_\phi(r_t|z_{t+1}, \hat{e}_{t+1}) p^i_\phi(i_{t+1}|z_{t+1}, \hat{e}_{t+1})], && (9) \end{aligned}$$

where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$.

To learn the parameters ϕ of this model $p_\phi(r_t, i_{t+1}|h_t, a_t)$, we may want to minimize the KL divergence from the true distribution, which is known to be equivalent to maximizing the log-likelihood of samples from the true distribution. Unfortunately, the log-likelihood $\log p_\phi(r_t, i_{t+1}|h_t, a_t)$ of such a model is not easy to estimate. Inspired by the evidence lower bound (ELBO) in variational inference, the Informed Dreamer proposes instead to maximize the informed ELBO,

$$\begin{aligned} &\widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t) \\ &= \mathop{\mathbb{E}}_{\tilde{O}(o_{t+1}|i_{t+1})} \Bigg[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \Bigg[\log p^r_\phi(r_t|z_{t+1}, e_{t+1}) + \log p^i_\phi(i_{t+1}|z_{t+1}, e_{t+1}) \\ &\qquad - \mathrm{KL}_{e'_{t+1}}(q^e_\phi(e'_{t+1}|z_{t+1}, o_{t+1}) \parallel p^{\hat{e}}_\phi(e'_{t+1}|z_{t+1})) \Bigg] \Bigg], \end{aligned} \tag{10}$$

where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$ and $q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1}) = q_\phi(e_{0:t+1}|h_{t+1})$ is the variational distribution, following equation (6). As proven in Subsection B.1, the informed ELBO is a valid lower bound on the log-likelihood: $\widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t) \leq \log p_\phi(r_t, i_{t+1}|h_t, a_t)$. The resulting world modeling objective is:

$$\max_\phi \mathop{\mathbb{E}}_{p(h_t, a_t, r_t, i_{t+1})} [\widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t)]. \tag{11}$$

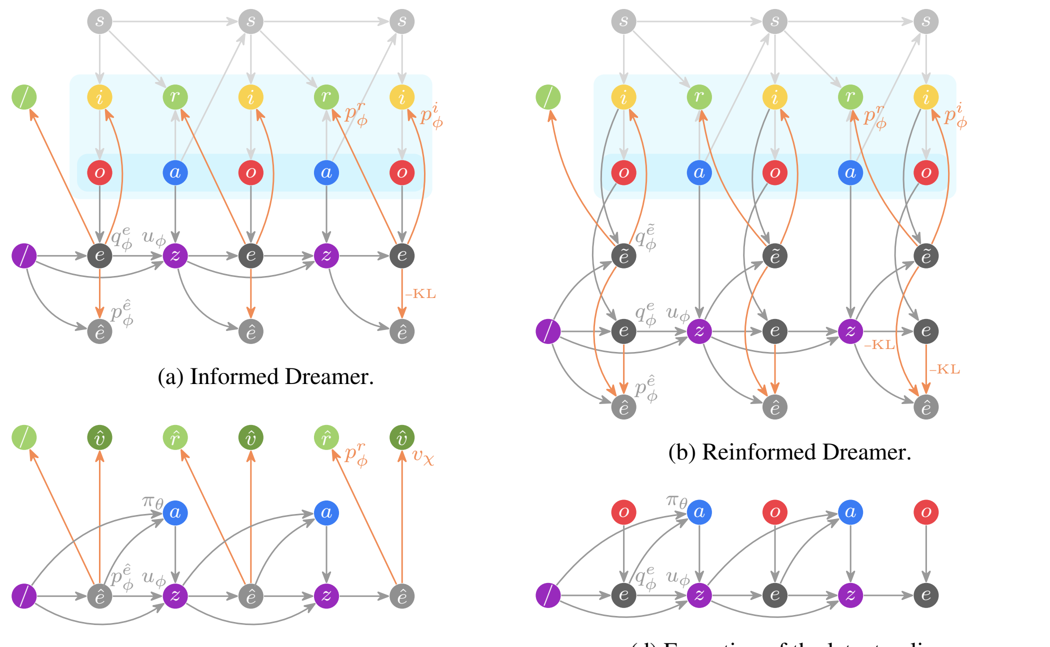
The world model architecture and its learning objective are illustrated in Figure 2a. Note that it also uses symlog predictions, a discrete latent space with straight-through gradients through categorical samples, and KL balancing with free bits as detailed in Appendix D.

Latent Policy. By selecting a latent policy $\pi_\theta(a_t|z_t, e_t)$, it is possible to generate sequences of interactions without reconstructing the observation nor the information. Indeed, starting from the initial state $z_0 = 0$, we sample the initial latent variable $e_0 \sim p^{\hat{e}}_\phi(e_0|z_0)$. Then, for any $t \geq 0$, we select the action $a_t \sim \pi_\theta(a_t|z_t, e_t)$, update the hidden state to $z_{t+1} = u_\phi(z_t, e_t, a_t)$, and sample the next latent $e_{t+1} \sim p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})$ from which we obtain the immediate reward $r_t \sim p^r_\phi(r_t|z_{t+1}, e_{t+1})$. In practice, the policy is optimized to maximize the imagined return by an on-policy actor-critic algorithm with a latent critic $v_\chi(z_t, e_t)$, using policy gradient ascent based on the latent trajectories. The latent imagination procedure is illustrated in Figure 2c.

While the latent variables are available during imagination, only the observations are available at execution. As a result, the latent policy $\pi_\theta(a_t|z_t, e_t)$ cannot be deployed directly. Instead, we infer the latent variables using the encoder $q_\phi(e_{0:t}|h_t) = \prod_{k=0}^{t} q^e_\phi(e_k|z_k, o_k)$ where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k < t$, and deploy the history-dependent policy $\eta_{\phi,\theta}(a_t|h_t)$, given by:

$$\eta_{\phi,\theta}(a_t|h_t) = \mathop{\mathbb{E}}_{q_\phi(e_{0:t}|h_t)} [\pi_\theta(a_t|z_t, e_t)]. \tag{12}$$

The policy execution is illustrated in Figure 2d.



(a) Informed Dreamer.

(b) Reinformed Dreamer.

(c) Latent imagination.

(d) Execution of the latent policy.

Figure 2: Training, imagination and execution of the Informed Dreamer and Reinformed Dreamer. In all figures, the dependence of the decoders and latent critic on the last hidden state is omitted.

3.2 Looseness of the Informed Evidence Lower Bound

When deriving the informed ELBO in equation (10) based on an observational encoder $q_\phi(e_k|z_k, o_k)$, we obtain a loose lower bound on the log-likelihood $\log p_\phi(r_t, i_{t+1}|h_t, a_t)$. Indeed, we have:

$$\begin{aligned}\widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t) &= \log p_\phi(r_t, i_{t+1}|h_t, a_t) \\ &\quad - \mathop{\mathbb{E}}_{\tilde{O}(o_{t+1}|i_{t+1})}\left[\mathrm{KL}_{e_{0:t+1}}(q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1}) \parallel p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1}))\right],\end{aligned} \tag{13}$$

where $q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})$ is the variational distribution and $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$ is the true posterior of the LVM defined by equation (9). The proof is given in Subsection B.2.

The KL term cannot be made arbitrarily small in general. First, the recurrent factorization of the variational distribution $q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})$ using the encoder $q^e_\phi(e_k|z_k, o_k)$ assumes that e_k is conditionally independent of the future given h_k, for all $k \leq t$. It is not true in general that the true posterior $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$ of the LVM has such conditional independence. This is known as the conditioning gap, which is inherent to the RSSM architecture (Becker & Neumann, 2022).

In addition, the variational distribution $q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})$ is conditioned on less information than the true posterior $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$, as it omits the reward r_t and the next information i_{t+1}. It is not true in general that the true posterior $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$ is independent of the reward r_t and next information i_{t+1} given the observation o_{t+1}, along with the history h_t and action a_t. We call this the information gap. As a consequence, the informed world model implemented by the LVM at equation (9) may be optimized through a loose lower bound, which can hinder learning.

In the particular case where the observation o_{t+1} does not contain any information about the information i_{t+1} (and the reward r_t), the ELBO will necessarily remain loose, and the distribution will be poorly approximated. Without any information about the information i_{t+1} (and the reward r_t) coming from the encoder, the learned decoder distribution will result in the best decoder fit, a problem akin to posterior collapse (He et al., 2019). In practice, the decoder class is a Gaussian of fixed variance, such that the best fit only encodes the conditional mean. Note that it is not a problem for the reward distribution, as we target the average return.

In the particular case where the observation o_{t+1} contains all information about the information i_{t+1} (and the reward r_t), the ELBO can become tight, up to the conditional gap of Becker & Neumann (2022), and the distribution may be correctly approximated. Since the informed ELBO can only be tight when the observation o_{t+1} contains all information about the information i_{t+1} (and the reward r_t), we conclude that the Informed Dreamer does not benefit from getting more information about the state s_{t+1} in the information i_{t+1} than in the observation o_{t+1}. Instead, we hypothesize that it may benefit from having less irrelevant information in the information i_{t+1} than in the observation o_{t+1}.

4 Reinformed Dreamer

To reduce the tightness gap of equation (13), we introduce the Reinformed Dreamer that has a privileged encoder to provide a tighter ELBO, and an unprivileged encoder trained by a latent guidance mechanism. The privileged encoder $q_\phi^{\tilde{e}}(e_{t+1}|z_{t+1}, i_{t+1})$ is conditioned on the next information to reduce the information gap. Note that it is not conditioned on the reward, as learning the conditional mean of the reward is sufficient for optimal control. To deploy the policy, we also keep the encoder $q_\phi^e(e_k|z_k, o_k)$ for $k \leq t$, which we call the unprivileged encoder. The resulting reinformed RSSM is:

$$e_k \sim q_\phi^e(e_k|z_k, o_k),\ k \leq t, \quad \text{(unprivileged encoder, 14)}$$
$$z_{k+1} = u_\phi(z_k, e_k, a_k),\ k \leq t, \quad \text{(recurrence, 15)}$$
$$\tilde{e}_{t+1} \sim q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1}), \quad \text{(privileged encoder, 16)}$$
$$\hat{e}_{t+1} \sim p_\phi^{\hat{e}}(\hat{e}_{t+1}|z_{t+1}), \quad \text{(prior, 17)}$$
$$\hat{r}_t \sim p_\phi^r(\hat{r}_t|z_{t+1}, \hat{e}_{t+1}), \quad \text{(reward decoder, 18)}$$
$$\hat{i}_{t+1} \sim p_\phi^i(\hat{i}_{t+1}|z_{t+1}, \hat{e}_{t+1}). \quad \text{(information decoder, 19)}$$

This RSSM implements the same LVM as that of the Informed Dreamer at equation (9), where the privileged encoder is not used. However, to train this LVM, we propose to use the privileged encoder to tighten the ELBO, resulting in the reinformed ELBO:

$$\mathrm{ELBO}_\phi(r_t, i_{t+1}|h_t, a_t) = \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})}\Big[\log p_\phi^r(r_t|z_{t+1}, \tilde{e}_{t+1}) + \log p_\phi^i(i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) - \mathrm{KL}_{\tilde{e}'_{t+1}}(q_\phi^{\tilde{e}}(\tilde{e}'_{t+1}|z_{t+1}, i_{t+1}) \parallel p_\phi^{\hat{e}}(\tilde{e}'_{t+1}|z_{t+1}))\Big], \quad (20)$$

where $q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1}) = q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})$ is the variational distribution, with $q_\phi(e_{0:t}|h_t)$ given by equation (6), and $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$. The proof is given in Subsection C.1. Interestingly, this lower bound is tight when $\mathrm{KL}_{e_{0:t}, \tilde{e}_{t+1}}(q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1}) \parallel p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, r_t, i_{t+1})) = 0$, where $p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, r_t, i_{t+1})$ is the true posterior of the LVM defined by equation (9). The proof is in Subsection C.2. More precisely, because the privileged variational distribution is conditioned on more information than the unprivileged variational distribution of the Informed Dreamer, the corresponding gap will be smaller than the tightness gap identified in equation (13). There may however remain a conditioning gap resulting from the recurrence of the variational distribution, as identified by Becker & Neumann (2022). To also train the unprivileged encoder (14) needed at execution as in the Informed Dreamer, we use the following learning objective:

$$\mathrm{ELBO}_\phi^+(r_t, i_{t+1}|h_t, a_t) = \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})}\Big[\log p_\phi^r(r_t|z_{t+1}, \tilde{e}_{t+1}) + \log p_\phi^i(i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) - \mathrm{KL}_{\tilde{e}'_{t+1}}(q_\phi^{\tilde{e}}(\tilde{e}'_{t+1}|z_{t+1}, i_{t+1}) \parallel p_\phi^{\hat{e}}(\tilde{e}'_{t+1}|z_{t+1})) - \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})}\Big[\mathrm{KL}_{e'_{t+1}}(q_\phi^e(e'_{t+1}|z_{t+1}, o_{t+1}) \parallel p_\phi^{\hat{e}}(e'_{t+1}|z_{t+1}))\Big]\Big]. \quad (21)$$

As seen in equation (21), the unprivileged encoder $q^{e}_{\phi}(e_t|z_t, o_t)$ is trained to match the prior $p^{\hat{e}}_{\phi}(e_t|z_t)$, and vice versa. The prior $p^{\hat{e}}_{\phi}(e_t|z_t)$ itself is trained to match the privileged encoder $q^{\tilde{e}}_{\phi}(e_t|z_t, i_t)$, and vice versa. We refer to this overall learning process for the prior, unprivileged encoder and privileged encoder as the latent guidance mechanism.

Finally, although not necessary, we propose to further increase the supervision signal on the unprivileged encoder by adding a reconstruction loss using the same decoders as for the privileged encoder. The final world modeling objective for the Reinformed Dreamer algorithm is:

$$
\begin{aligned}
\mathrm{ELBO}^{++}_{\phi}(r_t, i_{t+1}|h_t, a_t) = \underset{q_{\phi}(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})}{\mathbb{E}} \Bigg[& \log p^{r}_{\phi}(r_t|z_{t+1}, \tilde{e}_{t+1}) + \log p^{i}_{\phi}(i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) \\
& - \mathrm{KL}_{\tilde{e}'_{t+1}}(q^{\tilde{e}}_{\phi}(\tilde{e}'_{t+1}|z_{t+1}, i_{t+1}) \parallel p^{\hat{e}}_{\phi}(\tilde{e}'_{t+1}|z_{t+1})) \\
+ \underset{\widetilde{O}(o_{t+1}|i_{t+1})}{\mathbb{E}} \Bigg[\underset{q^{e}_{\phi}(e_{t+1}|z_{t+1}, o_{t+1})}{\mathbb{E}} & \Big[\log p^{r}_{\phi}(r_t|z_{t+1}, e_{t+1}) + \log p^{i}_{\phi}(i_{t+1}|z_{t+1}, e_{t+1}) \Big] \\
& - \mathrm{KL}_{e'_{t+1}}(q^{e}_{\phi}(e'_{t+1}|z_{t+1}, o_{t+1}) \parallel p^{\hat{e}}_{\phi}(e'_{t+1}|z_{t+1})) \Bigg] \Bigg].
\end{aligned} \tag{22}
$$

The architecture and learning objectives of the Reinformed Dreamer are illustrated in Figure 2b. As the Informed Dreamer, it uses symlog predictions, a discrete latent space with the Gumbel softmax reparametrization trick, and KL balancing with free bits as detailed in Appendix D. The imagination and execution procedures are unchanged and are illustrated in Figure 2c and Figure 2d, respectively.

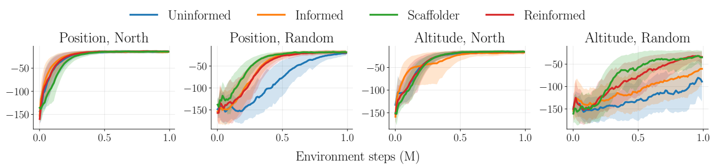
5 Experiments

We evaluate the Reinformed Dreamer on three suites of informed POMDPs, where privileged information is available during training. First, the Varying Mountain Hike environments (Igl et al., 2018; Lambrechts et al., 2024) are navigation tasks where the observation is a noisy measurement of the altitude, while the privileged information is the exact position and orientation of the agent. Second, the Pop Gym suite (Morad et al., 2023) is a suite of partially observable environments in which the privileged information is the full state. Finally, the Velocity Control suite (Tassa et al., 2018) is a version of the DeepMind Control Suite where only the velocity of the joints is observable, and the additional information is the exact position of the joints. We compare against the Uninformed Dreamer, the Informed Dreamer, and the Scaffolder algorithm, using the same hyperparameters as in the original implementations, which are all based on the DreamerV3 codebase. All results are averaged over five seeds. Other implementation details are reported in Appendix D.

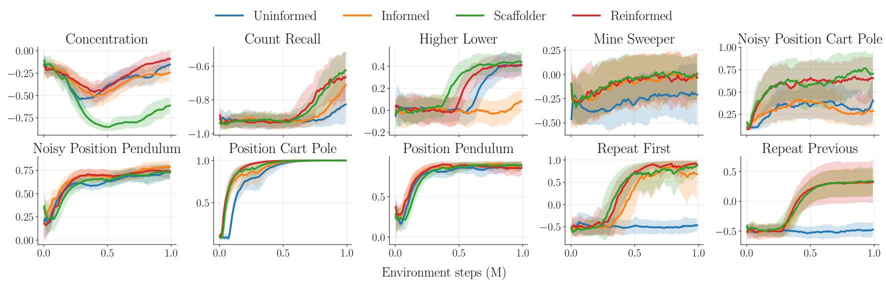
As shown in Figure 3, the Reinformed Dreamer outperforms or matches the performance of the Uninformed Dreamer and Informed Dreamer in all but four environments. The Reinformed Dreamer also outperforms or matches the performance of the Scaffolder in most environments, except two Velocity Control environments, and two Varying Mountain Hike environments. We consider these to be promising results that motivate further work on more challenging tasks. As shown in Table 1, the Reinformed Dreamer is about 20% slower to train than the Uninformed Dreamer and Informed Dreamer, versus approximately 80% to 120% for the Scaffolder and its two world models.

Table 1: Average runtime for training over one million steps across each suite for each algorithm.

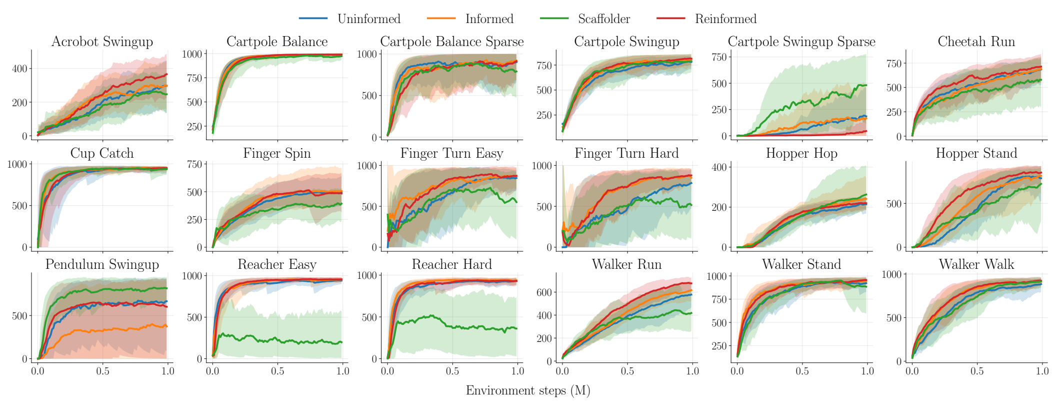
Algorithm	Varying Mountain Hike		Pop Gym		Velocity Control	
Uninformed Dreamer	2570.3 s	100.0 %	2274.7 s	100.0 %	36 794.5 s	100.0 %
Informed Dreamer	2537.5 s	98.7 %	2264.1 s	99.5 %	36 896.7 s	100.3 %
Reinformed Dreamer	3096.1 s	120.5 %	2782.2 s	122.3 %	43 298.0 s	117.7 %
Scaffolder	4686.4 s	182.3 %	4275.2 s	187.9 %	79 689.1 s	216.6 %



(a) Varying Mountain Hike Environments



(b) Pop Gym Environments



(c) Velocity Control environments

Figure 3: Evolution of the cumulative reward over one million training steps across all environments. For each method, we report the minimum, mean and maximum performance over five training seeds.

6 Conclusion

In this work, we improved asymmetric model-based RL under partial observability by studying the effect of asymmetric representation learning on the representation of the observation and on the representation of the additional information. First, we showed that the Informed Dreamer uses a loose lower bound on the log-likelihood of the privileged information, and argued that it hinders the learned representation of the additional information in practice. Second, we proposed the Reinformed Dreamer that provides a lower bound on the log-likelihood of the privileged information that has a smaller tightness gap. This tighter lower bound is obtained through a privileged encoder, which is used to train the unprivileged encoder through a latent guidance mechanism. Third, we demonstrated empirically that this provides an improved learning speed and final performance on most considered partially observable environments compared to the Uninformed Dreamer and Informed Dreamer, and that it matches the performance of the Scaffolder with a single world model.

Acknowledgments

Gaspard Lambrechts is a *Marie Skłodowska-Curie Actions* (MSCA) postdoctoral fellow at McGill University in Canada, funded by *Horizon Europe* under grant 101284668. Gaspard Lambrechts carried out part of this work as a *Fund for Scientific Research* (FNRS) postdoctoral fellow at the University of Liège in Belgium, funded by the *Wallonia-Brussels Federation*. Gaspard Lambrechts also gratefully acknowledges the financial support of *Wallonia-Brussels International* for his *World Excellence Fellowship*. Adrien Bolland is a *Fund for Scientific Research* (FNRS) postdoctoral fellow at the University of Liège in Belgium, funded by the *Wallonia-Brussels Federation*. Daniel Ebi acknowledges financial support by the *German Research Foundation* (DFG) as part of the Research Training Group GRK 2153. Computational resources have been provided by the *Consortium des Équipements de Calcul Intensif* (CÉCI), funded by the *Fund for Scientific Research* (FNRS) under grant 2502011 and by the *Walloon Region*, including the Tier-1 supercomputer of the *Wallonia-Brussels Federation*, infrastructure funded by the *Walloon Region* under grant 1117545.

References

Christopher Amato. An Introduction to Centralized Training for Decentralized Execution in Cooperative Multi-Agent Reinforcement Learning. *arXiv:2409.03052*, 2024.

Karl Johan Åström. Optimal Control of Markov Processes with Incomplete State Information. *Journal of Mathematical Analysis and Applications*, 1965.

Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo Perez, and Diederik M Roijers. The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models. *International Conference on Learning Representations*, 2024.

Andrea Baisero and Christopher Amato. Unbiased Asymmetric Reinforcement Learning under Partial Observability. *International Conference on Autonomous Agents and Multiagent Systems*, 2022.

Bram Bakker. Reinforcement Learning with Long Short-Term Memory. *Advances in Neural Information Processing Systems*, 2001.

Philipp Becker and Gerhard Neumann. On Uncertainty in Deep State Space Models for Model-Based Reinforcement Learning. *Transactions on Machine Learning Research*, 2022.

José M Bernardo and Adrian FM Smith. *Bayesian Theory*. John Wiley & Sons, 2009.

Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*. Athena Scientific, 2012.

Lars Buesing, Théophane Weber, Sébastien Racanière, S. M. Ali Eslami, Danilo Jimenez Rezende, David P. Reichert, Fabio Viola, Frederic Besse, Karol Gregor, Demis Hassabis, and Daan Wierstra. Learning and Querying Fast Generative Models for Reinforcement Learning. *arXiv:1802.03006*, 2018.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable Partially Observable Reinforcement Learning with Privileged Information. *Advances in Neural Information Processing Systems*, 2024.

Semih Cayci and Atilla Eryilmaz. Recurrent Natural Policy Gradient for POMDPs. *ICML Workshop on the Foundations of Reinforcement Learning and Control*, 2024.

Semih Cayci, Niao He, and R Srikant. Finite-Time Analysis of Natural Actor-Critic for POMDPs. *SIAM Journal on Mathematics of Data Science*, 2024.

Sanjiban Choudhury, Mohak Bhardwaj, Sankalp Arora, Ashish Kapoor, Gireeja Ranade, Sebastian Scherer, and Debadeepta Dey. Data-Driven Planning via Imitation Learning. *The International Journal of Robotics Research*, 2018.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan D. Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, J-M. Moret, Seb Noury, Federico Pesamosca, D. Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, B. Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin A. Riedmiller. Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning. *Nature*, 2022.

Peter Dürr, Mireille El Gheche, Guilherme Jorge Maeda, Nobuhiko Mukai, Naoya Takahashi, Stefan Heusser, Hamdi Sahloul, Yamen Saraiji, Pavel Adodin, Yin Bi, et al. Outplaying elite table tennis players with an autonomous robot. *Nature*, 2026.

Daniel Ebi, Gaspard Lambrechts, Damien Ernst, and Klemens Böhm. Informed Asymmetric Actor-Critic: Theoretical Insights and Open Questions. *European Workshop on Reinforcement Learning*, 2025.

Daniel Ebi, Damien Ernst, Klemens Böhm, and Gaspard Lambrechts. Informed Asymmetric Actor-Critic: Leveraging Privileged Signals Beyond Full-State Access. *International Conference on Machine Learning*, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual Multi-Agent Policy Gradients. *AAAI Conference on Artificial Intelligence*, 2018.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping Belief States with Generative Environment Models for RL. *Advances in Neural Information Processing Systems*, 2019.

Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A Pires, and Ré Munosmi. Neural Predictive Belief Representations. *arXiv:1811.06407*, 2018.

Zhaohan Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-Bastien Grill, Florent Altché, Rémi Munos, and Mohammad Gheshlaghi Azar. Bootstrap Latent-Predictive Representations for Multitask Reinforcement Learning. *International Conference on Machine Learning*, 2020.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. *Advances in Neural Information Processing Systems*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with Discrete World Models. *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Control Tasks through World Models. *Nature*, 2025.

Dongqi Han, Kenji Doya, and Jun Tani. Variational Recurrent Models for Solving Partially Observable Control Tasks. *Internal Conference on Learning Representations*, 2019.

Matthew Hausknecht and Peter Stone. Deep Recurrent Q-learning for Partially Observable MDPs. *AAAI Fall Symposium Series*, 2015.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. 2019.

Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. Memory-Based Control with Recurrent Neural Networks. *arXiv:1512.04455*, 2015.

Yitian Hong, Yaochu Jin, and Yang Tang. Rethinking Individual Global Max in Cooperative Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 2022.

Edward S Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged Sensing Scaffolds Reinforcement Learning. *International Conference on Learning Representations*, 2024.

Edward S Hu, Jie Wang, Xingfang Yuan, Fiona Luo, Muyao Li, Gaspard Lambrechts, Oleh Rybkin, and Dinesh Jayaraman. Real-World Reinforcement Learning of Active Perception Behaviors. *Advances in Neural Information Processing Systems*, 2025.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep Variational Reinforcement Learning for POMDPs. *International Conference on Machine Learning*, 2018.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-Level Drone Racing using Deep Reinforcement Learning. *Nature*, 2023.

Gaspard Lambrechts. *Reinforcement Learning in Partially Observable Markov Decision Processes: Learning to Remember the Past by Learning to Predict the Future*. PhD thesis, University of Liège, 2025.

Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Recurrent Networks, Hidden States and Beliefs in Partially Observable Environments. *Transactions on Machine Learning Research*, 2022.

Gaspard Lambrechts, Adrien Bolland, and Damien Ernst. Informed POMDP: Leveraging Additional Information in Model-Based RL. *Reinforcement Learning Journal*, 2024.

Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A Theoretical Justification for Asymmetric Actor-Critic Algorithms. *International Conference on Machine Learning*, 2025.

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model. *Advances in Neural Information Processing Systems*, 2020.

Pascal Leroy, Damien Ernst, Pierre Geurts, Gilles Louppe, Jonathan Pisane, and Matthia Sabatelli. QVMix and QVMix-Max: Extending the Deep Quality-Value Family of Algorithms to Cooperative Multi-Agent Reinforcement Learning. *AAAI Workshop on Reinforcement Learning in Games*, 2021.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust Multi-Agent Reinforcement Learning via Minimax Deep Deterministic Policy Gradient. *AAAI Conference on Artificial Intelligence*, 2019.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in Neural Information Processing Systems*, 2017.

Xueguang Lyu and Yuchen Xiao. A Deeper Understanding of State-Based Critics in Multi-Agent Reinforcement Learning. *AAAI Conference on Artificial Intelligence*, 2022.

John McCarthy. What is Artificial Intelligence, 1998.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. POP-Gym: Benchmarking Partially Observable Reinforcement Learning. *International Conference on Learning Representations*, 2023.

Tianwei Ni, Benjamin Eysenbach, Erfan SeyedSalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging State and History Representations: Understanding Self-Predictive RL. *International Conference on Learning Representations*, 2024.

Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. Optimal and Approximate Q-Value Functions for Decentralized POMDPs. *Journal of Artificial Intelligence Research*, 2008.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric Actor Critic for Image-Based Robot Learning. *Robotics: Science and Systems*, 2018.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *International Conference on Machine Learning*, 2018.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMix: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 2020.

Stuart J Russell. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2010.

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering Memory Tasks with World Models. *International Conference on Learning Representations*, 2024.

Amit Sinha and Aditya Mahajan. Asymmetric Actor-Critic with Approximate Information State. *IEEE Conference on Decision and Control*, 2023.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. *International Conference on Machine Learning*, 2019.

Charlotte T Striebel. Sufficient Statistics in the Optimum Control of Stochastic Systems. *Journal of Mathematical Analysis and Applications*, 1965.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate Information State for Approximate Planning and Reinforcement Learning in Partially Observed Systems. *Journal of Machine Learning Research*, 2022.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind Control Suite. *arXiv:1801.00690*, 2018.

Vladimir Vapnik and Akshay Vashist. A New Learning Paradigm: Learning using Privileged Information. *Neural Networks*, 2009.

Miguel Vasco, Takuma Seno, Kenta Kawamoto, Kaushik Subramanian, Peter R Wurman, and Peter Stone. A Super-Human Vision-Based Reinforcement Learning Agent for Autonomous Racing in Gran Turismo. *Reinforcement Learning Journal*, 2024.

Andrew Wang, Andrew C Li, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. Learning Belief Representations for Partially Observable Deep RL. *International Conference on Machine Learning*, 2023.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: Duplex Dueling Multi-Agent Q-Learning. *International Conference on Learning Representations*, 2021.

Rose E Wang, Michael Everett, and Jonathan P How. R-MADDPG for Partially Observable Environments and Limited Communication. *arXiv:2002.06684*, 2020.

Andrew Warrington, Jonathan W Lavington, Adam Scibior, Mark Schmidt, and Frank Wood. Robust Asymmetric Learning in POMDPs. *International Conference on Machine Learning*, 2021.

Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Solving Deep Memory POMDPs with Recurrent Policy Gradients. *International Conference on Artificial Neural Networks*, 2007.

Marvin Zhang, Zoe McCarthy, Chelsea Finn, Sergey Levine, and Pieter Abbeel. Learning Deep Neural Network Policies with Continuous Memory States. *IEEE International Conference on Robotics and Automation*, 2016.

Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On Improving Deep Reinforcement Learning for POMDPs. *arXiv:1704.07978*, 2017.

A Extended Related Work

In this section, we give an overview of RL under partial observability and asymmetric observability, present the theory of asymmetric RL, and discuss the related literature in multiagent RL.

A.1 Reinforcement learning under partial observability

Many RL approaches for partially observable environments have relied on learning a history-dependent policy, usually implemented using a recurrent neural network. These policies are typically trained with gradient-based RL algorithms using backpropagation through time (Bakker, 2001; Wierstra et al., 2007; Hausknecht & Stone, 2015; Heess et al., 2015; Zhang et al., 2016; Zhu et al., 2017). In this case, the hidden state of the recurrent neural network is a statistic of the history, which should ideally summarize all relevant information from the history to act optimally in the future (Lambrechts et al., 2022). Such a statistic of the history is called a sufficient statistic for optimal control (Striebel, 1965) or an information state (Bertsekas, 2012). Although these approaches theoretically allow implicit learning of a sufficient statistic, sufficient statistics can also be learned explicitly. Notably, many works (Igl et al., 2018; Buesing et al., 2018; Guo et al., 2018; Gregor et al., 2019; Han et al., 2019; Guo et al., 2020; Lee et al., 2020; Hafner et al., 2019; 2020) focused on learning a recurrent statistic that encodes the reward and next observation distribution given the action, a property known as predictive sufficiency (Bernardo & Smith, 2009). A recurrent and predictive statistic is indeed proven to be sufficient for optimal control (Subramanian et al., 2022). In practice, the sufficiency objective is usually optimized jointly with the RL objective.

A.2 Asymmetric reinforcement learning

As far as asymmetric RL is concerned, we identify four categories of algorithms: imitation learning, asymmetric critics, asymmetric world models, and asymmetric representation learning. Early approaches proposed to imitate a privileged policy conditioned on the state (Choudhury et al., 2018), or to use an asymmetric critic conditioned on the state (Pinto et al., 2018). These heuristic methods initially lacked a theoretical framework, and a recent line of work has focused on proposing theoretically grounded asymmetric learning objectives. First, imitation learning of a privileged policy was known to be suboptimal (Choudhury et al., 2018), and it was addressed by constraining the privileged policy so that its imitation results in an optimal policy for the partially observable environment (Warrington et al., 2021). Similarly, asymmetric actor-critic approaches were proven to provide biased gradients, and an unbiased actor-critic approach was proposed by introducing the history-state value function (Baisero & Amato, 2022). Many recent empirical successes of RL have relied on an asymmetric critic (Degrave et al., 2022; Kaufmann et al., 2023; Vasco et al., 2024; Hu et al., 2024; 2025; Dürr et al., 2026). In model-based RL, several works proposed world model objectives that are proven to provide sufficient statistics of the history for optimal control, by leveraging the state (Avalos et al., 2024) or arbitrary state information (Lambrechts et al., 2024). Finally, asymmetric representation learning approaches were proposed to learn sufficient statistics of the history for optimal control, from state samples (Wang et al., 2023; Sinha & Mahajan, 2023).

A.3 Theory of asymmetric learning

While a line of work has focused on proposing theoretically grounded asymmetric learning objectives (Warrington et al., 2021; Baisero & Amato, 2022; Lambrechts et al., 2024; Wang et al., 2023), in the sense that optimizing for these objectives provides optimal policies, a theoretical justification for why these asymmetric objectives help was still missing. A recent line of work has focused on finding a theoretical justification for asymmetric actor-critic algorithms. A theoretical justification was proposed by adapting a finite-time convergence analysis for actor-critic algorithms with fixed agent state and linear function approximator (Cayci et al., 2024) to the asymmetric setting (Lambrechts et al., 2025). The comparison of these finite-time bounds highlighted an advantage of asymmetric learning in the presence of aliasing in the agent state space. An alternative finite-time bound was provided for a tabular belief-weighted asymmetric actor-critic, highlighting again the gains from the privileged information (Cai et al., 2024). However, both these theoretical works assumed privileged access to the full state during training. Recent work proposed to extend the asymmetric actor-critic algorithm to arbitrary additional state information (Ebi et al., 2026). Then, by adapting a finite-time convergence proof of recurrent actor-critic (Cayci & Eryilmaz, 2024), several hypotheses for the benefits of asymmetric recurrent actor-critic are discussed (Ebi et al., 2025). As far as asymmetric world models and asymmetric representation learning are concerned, a detailed theoretical justification for their effectiveness is still missing.

A.4 Centralized training for decentralized execution

In multiagent RL, exploiting additional information available during training was extensively studied under the centralized training for decentralized execution (CTDE) framework (Oliehoek et al., 2008; Amato, 2024). It is also worth noting that CTDE was definitely an inspiration for many asymmetric RL methods for partial observability. In CTDE, it is assumed that the histories of all agents, or even the environment state, are available to all agents during training. To exploit this additional information, several asymmetric actor-critic approaches have been developed by leveraging an asymmetric critic conditioned on all histories, including COMA (Foerster et al., 2018), MADDPG (Lowe et al., 2017), M3DDPG (Li et al., 2019) and R-MADDPG (Wang et al., 2020). While efficient in practice, Lyu & Xiao (2022) showed that these asymmetric actor-critic approaches provide biased gradient estimates, which generalizes results developed in the asymmetric learning literature (Baisero & Amato, 2022) to the multi-agent setting. In the cooperative CTDE setting, another line of work focuses on value decomposition to learn a utility function for each agent, including QMIX (Rashid et al., 2018), QVMix (Leroy et al., 2021) and QPLEX (Wang et al., 2021). These approaches use the additional information to modulate the contribution of each utility function in the global value function, while ensuring that maximizing the local utility functions also maximize the global value function, a property known as individual global maximum (IGM). Other methods relax this IGM requirement but still condition the value function on all histories, including QTRAN (Son et al., 2019) and WQMix (Rashid et al., 2020). Hong et al. (2022) established that the IGM decomposition is not attainable in the general case.

B Informed Evidence Lower Bound

In Subsection B.1, we introduce the informed ELBO as a lower bound on the log-likelihood. In Subsection B.2, we show that this lower bound is loose in general.

B.1 Informed Evidence Lower Bound

First, let us recall the prior distribution and the variational distribution of the informed RSSM:

$$q_\phi(e_{0:t}|h_t) = \textstyle\prod_{k=0}^{t} q^e_\phi(e_k|z_k, o_k), \tag{23}$$

$$p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t) = q_\phi(e_{0:t}|h_t) p^{\hat{e}}_\phi(\hat{e}_{t+1}|z_{t+1}), \tag{24}$$

$$q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1}) = q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1}) = q_\phi(e_{0:t+1}|h_{t+1}), \tag{25}$$

where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$. Starting from the log-likelihood $\log p_\phi(r_t, i_{t+1}|h_t, a_t)$ of the LVM at equation (9), we have:

$$\begin{aligned}
&\log p_\phi(r_t, i_{t+1}|h_t, a_t) \\
&\quad = \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} \left[p_\phi(r_t, i_{t+1}|z_{t+1}, \hat{e}_{t+1}) \right] && (26) \\
&\quad = \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \hat{e}_{t+1}) q^e_\phi(\hat{e}_{t+1}|z_{t+1}, o_{t+1})}{q^e_\phi(\hat{e}_{t+1}|z_{t+1}, o_{t+1})} \right] \right] && (27) \\
&\quad = \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{p_\phi(e_{0:t}, e_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})}{q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] && (28)
\end{aligned}$$

$$= \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{p_\phi(e_{0:t}, e_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})}{q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{29}$$

$$= \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{p_\phi(e_{0:t}, e_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})}{q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{30}$$

$$= \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p_\phi(e_{0:t}, e_{t+1}|h_t, a_t)}{q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{31}$$

$$= \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q_\phi(e_{0:t}|h_t) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{32}$$

$$= \log \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{33}$$

$$\geq \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{34}$$

$$\begin{aligned} = \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \Bigg[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \Bigg[&\log p^r_\phi(r_t|z_{t+1}, e_{t+1}) + \log p^i_\phi(i_{t+1}|z_{t+1}, e_{t+1}) \\ &- \mathrm{KL}_{e'_{t+1}}(q^e_\phi(e'_{t+1}|z_{t+1}, o_{t+1}) \parallel p^{\hat{e}}_\phi(e'_{t+1}|z_{t+1})) \Bigg] \Bigg] \end{aligned} \tag{35}$$

$$= \widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t). \tag{36}$$

B.2 Informed Tightness Gap

Starting from equation (34), we can develop the ELBO further. We have:

$$\widetilde{\mathrm{ELBO}}_\phi(r_t, i_{t+1}|h_t, a_t)$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{37}$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{38}$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q_\phi(e_{0:t}|h_t) p^{\hat{e}}_\phi(e_{t+1}|z_{t+1})}{q_\phi(e_{0:t}|h_t) q^e_\phi(e_{t+1}|z_{t+1}, o_{t+1})} \right] \right] \tag{39}$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p_\phi(e_{0:t+1}|h_t, a_t)}{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \right] \right] \tag{40}$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(e_{0:t+1}, r_t, i_{t+1}|h_t, a_t)}{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \right] \right] \tag{41}$$

$$= \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|h_t, a_t) p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})}{q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1})} \right] \right] \tag{42}$$

$$\begin{aligned} &= \log p_\phi(r_t, i_{t+1}|h_t, a_t) \\ &\quad - \mathop{\mathbb{E}}_{\widetilde{O}(o_{t+1}|i_{t+1})} \left[\mathrm{KL}_{e_{0:t+1}}(q_\phi(e_{0:t+1}|h_t, a_t, o_{t+1}) \parallel p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})) \right], \end{aligned} \tag{43}$$

where $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$ is the true posterior of the LVM.

C Reinformed Evidence Lower Bound

In Subsection C.1, we introduce the reinformed ELBO as a lower bound on the log-likelihood. In Subsection C.2, we characterize the tightness gap of this lower bound.

C.1 Reinformed Evidence Lower Bound

First, let us recall the prior distribution and the variational distribution of the reinformed RSSM:

$$q_\phi(e_{0:t}|h_t) = \prod_{k=0}^{t} q_\phi^e(e_k|z_k, o_k), \tag{44}$$

$$p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t) = q_\phi(e_{0:t}|h_t) p_\phi^{\hat{e}}(\hat{e}_{t+1}|z_{t+1}), \tag{45}$$

$$q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1}) = q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1}), \tag{46}$$

where $z_{k+1} = u_\phi(z_k, e_k, a_k)$ for $k \leq t$. Starting from the log-likelihood $\log p_\phi(r_t, i_{t+1}|h_t, a_t)$ of the LVM at equation (9), we have:

$$\begin{aligned}
&\log p_\phi(r_t, i_{t+1}|h_t, a_t) \\
&= \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} \left[p_\phi(r_t, i_{t+1}|z_{t+1}, \hat{e}_{t+1})\right] && (47) \\
&= \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \hat{e}_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \hat{e}_{t+1}) q_\phi^{\tilde{e}}(\hat{e}_{t+1}|z_{t+1}, i_{t+1})}{q_\phi^{\tilde{e}}(\hat{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (48) \\
&= \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}{q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (49) \\
&= \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}{q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (50) \\
&= \log \mathop{\mathbb{E}}_{p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t)} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})}{q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (51) \\
&= \log \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) p_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t)}{q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (52) \\
&= \log \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) q_\phi(e_{0:t}|h_t) p_\phi^{\hat{e}}(\tilde{e}_{t+1}|z_{t+1})}{q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (53) \\
&= \log \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})} \left[\frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) p_\phi^{\hat{e}}(\tilde{e}_{t+1}|z_{t+1})}{q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (54) \\
&\geq \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})} \left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) p_\phi^{\hat{e}}(\tilde{e}_{t+1}|z_{t+1})}{q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (55) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})} \Big[\log p_\phi^r(r_t|z_{t+1}, \tilde{e}_{t+1}) + \log p_\phi^i(i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) \\
&\qquad - \mathrm{KL}_{\tilde{e}'_{t+1}}(q_\phi^{\tilde{e}}(\tilde{e}'_{t+1}|z_{t+1}, i_{t+1}) \parallel p_\phi^{\hat{e}}(\tilde{e}'_{t+1}|z_{t+1}))\Big] && (56) \\
&= \mathrm{ELBO}_\phi(r_t, i_{t+1}|h_t, a_t). && (57)
\end{aligned}$$

C.2 Reinformed Tightness Gap

Starting from equation (55), we can develop the ELBO further. We have:

$$\begin{aligned}
&\mathrm{ELBO}_\phi(r_t, i_{t+1}|h_t, a_t) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, \tilde{e}_{t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, \tilde{e}_{t+1}) p_\phi^{\hat{e}}(\tilde{e}_{t+1}|z_{t+1})}{q_\phi^{\tilde{e}}(\tilde{e}_{t+1}|z_{t+1}, i_{t+1})}\right] && (58) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t}, e_{t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p_\phi^{\hat{e}}(e_{t+1}|z_{t+1})}{q_\phi^{\tilde{e}}(e_{t+1}|z_{t+1}, i_{t+1})}\right] && (59) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p_\phi^{\hat{e}}(e_{t+1}|z_{t+1})}{q_\phi^{\tilde{e}}(e_{t+1}|z_{t+1}, i_{t+1})}\right] && (60) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) q_\phi(e_{0:t}|h_t) p_\phi^{\hat{e}}(e_{t+1}|z_{t+1})}{q_\phi(e_{0:t}|h_t) q_\phi^{\tilde{e}}(e_{t+1}|z_{t+1}, i_{t+1})}\right] && (61) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|z_{t+1}, e_{t+1}) p_\phi(e_{0:t+1}|h_t, a_t)}{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\right] && (62) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(e_{0:t+1}, r_t, i_{t+1}|h_t, a_t)}{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\right] && (63) \\
&= \mathop{\mathbb{E}}_{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\left[\log \frac{p_\phi(r_t, i_{t+1}|h_t, a_t) p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})}{q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1})}\right] && (64) \\
&= \log p_\phi(r_t, i_{t+1}|h_t, a_t) \\
&\qquad - \mathrm{KL}_{e_{0:t+1}}(q_\phi(e_{0:t+1}|h_t, a_t, i_{t+1}) \parallel p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})), && (65)
\end{aligned}$$

where $p_\phi(e_{0:t+1}|h_t, a_t, r_t, i_{t+1})$ is the true posterior of the LVM.

D Experimental Details

For the Uninformed Dreamer (Hafner et al., 2025), Informed Dreamer (Lambrechts et al., 2024) and Scaffolder (Hu et al., 2024) algorithms, we use the original implementations and hyperparameters. Similarly, the Reinformed Dreamer algorithm is based on the original Informed Dreamer implementation and hyperparameters.

As far as the training objectives of the unprivileged encoder and privileged encoder are concerned, we use the same hyperparameters as the observational encoder in the Uninformed Dreamer and Informed Dreamer. More precisely, we use KL balancing with $\beta_{\text{dyn}} = 0.5$ and $\beta_{\text{rep}} = 0.1$ and free bits of 1 nat for both encoders. By denoting the stop-gradient operator with $\mathrm{sg}(\cdot)$, we have:

$$\begin{aligned}
\widehat{\mathrm{KL}}_{e_t}(q_\phi^{\tilde{e}}(e_t|z_t, i_t) \parallel p_\phi^{\hat{e}}(e_t|z_t)) &= \beta_{\text{dyn}} \max(1, \mathrm{KL}_{e_t}(\mathrm{sg}(q_\phi^{\tilde{e}}(e_t|z_t, i_t)) \parallel p_\phi^{\hat{e}}(e_t|z_t))) \\
&\quad + \beta_{\text{rep}} \max(1, \mathrm{KL}_{e_t}(q_\phi^{\tilde{e}}(e_t|z_t, i_t) \parallel \mathrm{sg}(p_\phi^{\hat{e}}(e_t|z_t)))) && (66) \\
\widehat{\mathrm{KL}}_{e_t}(q_\phi^{e}(e_t|z_t, o_t) \parallel p_\phi^{\hat{e}}(e_t|z_t)) &= \beta_{\text{dyn}} \max(1, \mathrm{KL}_{e_t}(\mathrm{sg}(q_\phi^{e}(e_t|z_t, o_t)) \parallel p_\phi^{\hat{e}}(e_t|z_t))) \\
&\quad + \beta_{\text{rep}} \max(1, \mathrm{KL}_{e_t}(q_\phi^{e}(e_t|z_t, o_t) \parallel \mathrm{sg}(p_\phi^{\hat{e}}(e_t|z_t)))). && (67)
\end{aligned}$$

In all environments, we use $i_t = (o_t, o_t^+)$ for the Reinformed Dreamer even if the additional information o_t^+ is known to be a Markovian state, as we found it slightly more effective.