In-Context Reinforcement Learning via Communicative World Models*

Fernando Martinez[1], **Tao Li**[2†], **Yingdong Lu**[3] and **Juntao Chen**[1]
[1]Department of Computer and Information Sciences, Fordham University
[2]Department of Systems Engineering, City University of Hong Kong
[3]IBM Research
{fmartinezlopez, jchen504}@fordham.edu, li.tao@cityu.edu.hk, yingdong@us.ibm.com

Abstract

Reinforcement learning (RL) agents often struggle to generalize to new tasks and contexts without updating their parameters, mainly because their learned representations and policies are overfit to the specifics of their training environments. To boost agents' in-context RL (ICRL) ability, this work formulates ICRL as a two-agent emergent communication problem and introduces CORAL (Communicative Representation for Adaptive RL), a framework that learns a transferable communicative context by functionally separating latent representation learning from control. In CORAL, an Information Agent (IA) is pre-trained as a world model on a diverse distribution of tasks. Its objective is not direct return maximization, but world modeling and distilling its understanding into concise messages. The emergent communication protocol is shaped by a novel Causal Influence Loss, which measures the effect that the message has on the next action. During deployment, the previously trained IA serves as a fixed contextualizer for a new Control Agent (CA), which learns to solve tasks by interpreting the provided communicative context. Our experiments demonstrate that this approach enables the CA to achieve significant gains in sample efficiency and successfully perform zero-shot adaptation with the help of pre-trained IA in diverse online and offline environments, validating the efficacy of learning a transferable communicative representation.

1 Introduction

Pursuing a generalist agent, which is capable of solving a wide range of tasks with minimal intervention, has long been considered one of the central challenges of reinforcement learning (RL) and artificial intelligence (AI) at large [Reed, 2022]. Most recently, substantial effort has been dedicated to two distinct research thrusts targeting RL generalization: in-context reinforcement learning (ICRL) [Laskin *et al.*, 2023] and world models (WM) [Schrittwieser *et al.*, 2020; Hafner *et al.*, 2025]. The key idea behind ICRL is to condition the RL policy on some context variables in addition to the state observations. The adaptation power originates from the agent's response to the emerging context extracted from the past interactions [Laskin *et al.*, 2023; Li *et al.*, 2025], which reveals task-related information that improves generalization [Li *et al.*, 2023]. The defining characteristic of ICRL is that the in-context policy improvement is purely context-driven and does not rely on policy model updates as in gradient-based meta RL [Finn *et al.*, 2017; Li *et al.*, 2024; Pan *et al.*, 2025].

However, most recent ICRL approaches condition on past trajectories using the transformer architecture [Vaswani *et al.*, 2017] but typically do not understand the underlying task dynamics. Consequently, their performance depends largely on the quality of offline datasets [Chen *et al.*, 2021], which limits generalization beyond the distribution of training contexts [Chen *et al.*, 2024]. In contrast, world models (WMs), typically generative models, aim to equip agents with a structured understanding of the environment dynamics and reward feedback, which allows agents to predict how the environment will evolve in response to their actions, even those rarely seen in the pre-training dataset [Hafner *et al.*, 2020; Hafner *et al.*, 2025].

From an ICRL perspective, the latent representation learned by the WM, which captures the dynamics and task reward, provides a context for RL agents when deployed in a variety of environments. Our intuition is that WMs are more suitable as contextualizers than vanilla, self-supervised trained transformers, leading to a subsequent reinforcement pre-training in ICRL with improved generalizability [Moeini *et al.*, 2025]. In summary, while ICRL facilitates context-driven adaptation, it often lacks understanding of the environment dynamics, making the generalization fragile. While WMs learn such dynamics, the learned representations, however, are often entangled with task-specific policy learning, which can lead to representations that are overly specialized and less transferable, and the objective mismatch between representation and policy learning [Eysenbach *et al.*, 2022].

To address WMs' limitation above and make it better suited for ICRL, we propose to separate the representation learning from the policy learning and introduce the Communicative Representation for Adaptive RL (CORAL) framework, illustrated in Fig. 1. CORAL functionally separates the role of

*Code available at https://github.com/fernando-ml/CORAL
†Corresponding author: li.tao@cityu.edu.hk

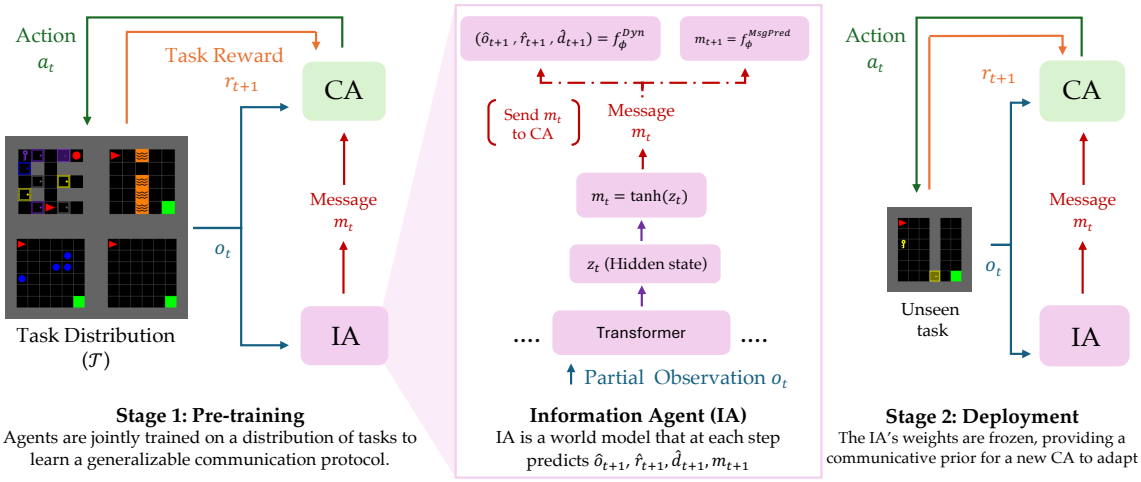

Figure 1: Overview of CORAL. **(Left) Pre-training:** An Information Agent (IA, ϕ) and Control Agent (CA, θ) are jointly trained on a task distribution $\mathcal{T}$. The IA processes observations o_t into messages m_t via self-supervised and communication objectives; the CA uses m_t to select actions. **(Right) Deployment:** The IA is frozen and provides a communicative prior to accelerate a new CA's adaptation to unseen tasks.

world modeling from reward maximization and formulates ICRL as a two-agent emergent communication problem [Lowe *et al.*, 2019; Zhu *et al.*, 2024], where a world model, called Information Agent (IA), aims to understand the environment dynamics and task reward, and communicate its understanding through latent representations to a Control Agent (CA), parameterized by a neural network policy model. CORAL's novelty lies in that IA, parameterized by a transformer, generates its messages considering three aspects: 1) dynamics awareness, messages encode the predictions of future consequences, 2) temporal coherence, consecutive messages in the same context must be alike to ensure consistency, and 3) in-context communication effectiveness, the control policy conditioned on messages yields higher return than without.

Our contributions are summarized below, and we postpone the discussion on related works to the end. 1) **Formulation**: We model a single-agent ICRL problem as a two-agent emergent communication where the world model (information agent) learns how to communicate the learned dynamics to the control agent. 2) **Hybrid Training**: since the information and control agents face distinct objectives, they are pre-trained differently. The transformer-based IA adopts a self-supervised training with three heads using three losses targeting dynamic awareness, temporal coherence, and communication effectiveness, while the CA adopts proximal policy optimization for reinforcement pre-training. 3) **Empirical Validation**: We provide extensive empirical validation in partially observable sparse-reward environments and continuous tasks, where CORAL improves sample efficiency most clearly in harder sparse-reward tasks, facilitates zero-shot generalization, and reaches higher performance.

2 Preliminaries

We formulate RL tasks as Partially Observable Markov Decision Processes (POMDP) [Kaelbling *et al.*, 1998]. The agent learns a policy π to select actions a_t based on a sequence of partial observations o_t, as the true environment state s_t is not directly accessible. The agent's objective is to maximize the expected discounted return $\mathbb{E}_\pi[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$, where r_t is the reward received after taking a_t and γ is the discount factor. We define the value function, parameterized by a neural network, as $V_\theta(o_t) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | o_t]$.

The emergent communication between two agents in a POMDP under a cheap-talk setting is, mathematically, a communication game [Crawford and Sobel, 1982; Li and Zhu, 2023]. In addition to the control agent whose setup is the same as above, an information agent adopts a communication policy $\pi^{\text{IA}}(\cdot|o_t)$, from which a message m_t is drawn. The communication is cheap because m_t does not incur cost to any agent nor influence state transition. In this case, the control agent's policy become $\pi^{\text{CA}}(\cdot|o_t, m_t)$, which still aims to maximizes the cumulative rewards, whereas the IA's return, fully determined by CA's actions, may share the same objective as CA's [Sukhbaatar *et al.*, 2016] or a totally different one [Crawford and Sobel, 1982].

3 Methodology

The ability to learn representations that generalize well and adapt quickly to new scenarios remains one of the main goals in reinforcement learning. End-to-end solutions in which the loss backpropagated through a single reward signal trains a single monolithic network can produce impressive results [Mnih *et al.*, 2015]. However, they often produce solutions that are heavily overfitted to the training task and struggle to generalize. Several methods have addressed this problem by decoupling the learning process: allowing agents to learn general-purpose, reusable representations based on self-supervised objectives [Stooke *et al.*, 2021; Schwarzer *et al.*, 2021]. These can learn strong features but may not always align those features with the downstream control task.

Overview. We introduce CORAL (Communicative Representation for Adaptive RL), a framework that presents a principled middle ground. CORAL frames the problem as one

of functional decomposition between representation and control. An Information Agent (IA) learns a dynamics model, while separately, a Control Agent (CA) maximizes reward. The learned protocol by IA can then be flexibly used as a strong contextual prior by CA to enable fast in-context adaptation.

CORAL builds on the asymmetric learning goals of the agents. The CA is a standard RL agent that is motivated by extrinsic task reward. In contrast, the IA is not trained on any direct return-maximization objective, and no policy gradient flows from the CA into the IA. Rather, the IA must learn a predictive model of the world's dynamics and compress this into a communicative representation. To ensure this separated representation remains relevant, the information transmitted by the IA is aligned with the CA's control policy through an influence-theoretic objective.

This alignment is fostered through a composite objective consisting of a self-supervised dynamics modeling loss and a causal influence objective. Constraining the IA to predict future observations, rewards, and terminations causes its internal state to represent physical truths about the environment. Meanwhile, its Causal Influence Loss incentivizes it to send specific messages. In particular, this loss causes the IA to send messages that result in a meaningful change to the CA's policy, but only when that change correlates with high-utility actions. Utility is calculated as a mixture of long-term Generalized Advantage Estimate rewards and the immediate change in the CA's own value prediction, forcing it to learn to communicate both immediately relevant and forward-looking information. Because these two properties are learned from functionally distinct principles, the resultant general representation serves as a powerful pre-trained communicative prior upon deployment.

3.1 CORAL Agents and Emergent Communication

CORAL is instantiated with two separate neural networks, which interact with one another at each timestep t. Each agent observes the same partial observation $o_t \in \mathbb{R}^{D_{\text{obs}}}$, but has distinct responsibilities. The Information Agent (IA), with parameters ϕ, aims to store past information and model the environment. We implement this agent with a Transformer-based architecture [Vaswani *et al.*, 2017] that takes in a context $c_t = (e_{t-L+1}, \ldots, e_t)$ of the L most recent embedded observations as input. At timestep t, we embed the observation o_t into e_t with a linear layer. The context c_t is stored in a sliding buffer and passed through multiple self-attention and feed-forward layers with residual connections. Since each token in the context attends to every other token, the contextualized sequence of output vectors encapsulates information about the recent history of e_t. Appendix G offers a gradient-dynamics interpretation of our transformer-based in-context learning and its relations with other in-context RL architectures.

We take the final output vector z_t, corresponding to the most recent observation token within c_t, as the IA's state representation. This allows the IA to use its observational history when choosing what information to convey to the CA. The emitted message from the IA is generated by another linear layer with a $\tanh$ activation: $m_t = \tanh(f_\phi^{\text{Msg}}(z_t))$. The $\tanh$ non-linearity is used to constrain the resulting message vector. The IA is not trained on any direct return-maximization objective; no policy gradient flows from the CA into the IA.

The Control Agent (CA) is parameterized by θ and is tasked with controlling actions. It operates as a standard RL agent whose policy $\pi_\theta(\cdot|o_t, m_t)$ and value function $V_\theta(o_t, m_t)$ are conditioned on both the observation and the message. While CORAL is algorithm-agnostic, we instantiate the CA with Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017] for our primary online adaptation experiments. Its objective is to maximize the task reward.

3.2 Pre-train the Communicative Representation

The agents' parameters ϕ, θ are optimized concurrently during pre-training. The learning objectives are constructed so as to encourage an IA that captures the world's dynamics and can communicate that knowledge, and a CA that can use that communication to solve the task.

The Information Agent as a Communicative World Model The IA is trained via a composite self-supervised loss function, $\mathcal{L}(\phi)$, that does not depend directly on the extrinsic task reward. It combines objectives for world dynamics modeling, message consistency, and communicative efficacy.

$$\mathcal{L}(\phi) = \lambda_{\text{Dyn}}\mathcal{L}_{\text{Dyn}}(\phi) + \lambda_{\text{Coh}}\mathcal{L}_{\text{Coh}}(\phi) + \lambda_{\text{Causal}}\mathcal{L}_{\text{Causal}}(\phi).$$

Dynamics Awareness Loss ($\mathcal{L}_{\text{Dyn}}$): This objective anchors the IA's representations by learning to anticipate the results of the CA's actions. From the message m_t and action a_t selected by the CA, the IA predicts the next observation $\hat{o}_{t+1}$, reward $\hat{r}_{t+1}$, and termination probability $\hat{d}_{t+1}$. The loss $\mathcal{L}_{\text{Dyn}}$ is given below, and see Appendix C for more details.

$$\mathbb{E}_t[\|\hat{o}_{t+1} - o_{t+1}\|^2 + (\hat{r}_{t+1} - r_{t+1})^2 + \text{BCE}(\hat{d}_{t+1}, d_{t+1})],$$

where predictions are generated by prediction heads (f_ϕ^{Dyn}) conditioned on the message m_t and action a_t: $(\hat{o}'_{t+1}, \hat{r}_{t+1}, \hat{d}_{t+1}) = f_\phi^{\text{Dyn}}(m_t, a_t)$.The first two terms in the loss use mean-squared error since the observation and reward predictions are continuous, whereas the last term uses binary cross-entropy (BCE) for probabilistic outputs.

Temporal Coherence Loss ($\mathcal{L}_{\text{Coh}}$): To promote temporal coherence in communication, the IA is also trained to predict the actual next-step message m_{t+1} based on the current message m_t through a specific prediction head (f_ϕ^{MsgPred}) with the loss given by $\mathcal{L}_{\text{Coh}} = \mathbb{E}_t\left[\|\hat{m}_{t+1} - m_{t+1}\|^2\right]$, where $\hat{m}_{t+1} = f_\phi^{\text{MsgPred}}(m_t, a_t)$. We encourage the message to be a compact representation of the state from which future states (and thus future messages) can be inferred.

Causal Influence Loss ($\mathcal{L}_{\text{Causal}}$): To shape the communication to be effective from the CA's perspective, we introduce an information-theoretic objective inspired by the work on measuring causal influence in emergent communication [Lowe *et al.*, 2019]. This loss encourages the IA to produce messages that cause a beneficial and decisive shift in the CA's policy. We quantify this per-step shift using a metric we term the Instantaneous Causal Effect (ICE), defined as the Reverse KL Divergence between the CA's policy with and without the message:

$$\text{ICE}_t = D_{\text{KL}}\Big(\pi_\theta(\cdot|o_t, \mathbf{0}) \parallel \pi_\theta(\cdot|o_t, m_t)\Big). \qquad \text{(ICE)}$$

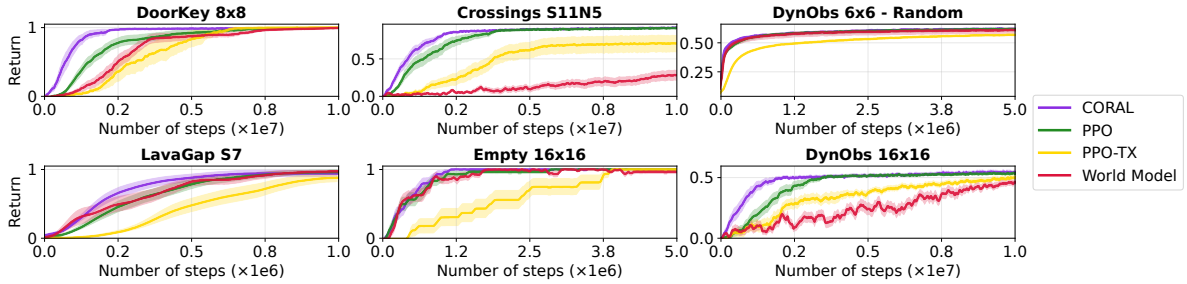

Figure 2: In-context adaptation with a pre-trained Information Agent. Learning curves show the mean episodic return ($\pm 95\%$ CI) across 30 seeds for a randomly initialized Control Agent paired with a pre-trained, frozen CORAL IA. CORAL demonstrates higher sample efficiency and asymptotic performance compared to a standard PPO, PPO-Transformer, and an equivalent World Model across unseen MiniGrid environments.

Here, $\pi_\theta(\cdot|o_t, \mathbf{0})$ is conditioned on a zero-vector message, representing a non-informative communicative input.

To encourage helpful messages from IA that yield high-return actions, we multiply the ICE by a hybrid utility signal $\mathcal{U}_t$. The signal combines two sources of utility: The first is the Advantage estimate A_t (e.g., GAE [Schulman *et al.*, 2015]), which provides a low-variance estimate of the long-term benefit of taking an action. It is calculated as the exponentially-weighted average of temporal difference errors: $A_t = \sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k}$, where $\delta_t = r_{t+1} + \gamma V_{t+1} - V_t$. The second source is the immediate change in the CA's own value estimate, $\Delta V_t = V_\theta(o_t, m_t) - V_\theta(o_t, \mathbf{0})$. To ensure that both signals contribute on a comparable scale regardless of reward magnitude or the stage of training, we apply z-score normalization, a standard technique for stabilizing policy gradient updates [Schulman *et al.*, 2017; Ioffe and Szegedy, 2015]. These normalized signals are then combined and clipped to form the final positive utility signal $\mathcal{U}_t = \max\{0, \alpha \cdot \text{norm}(\Delta V_t) + (1 - \alpha) \cdot \text{norm}(A_t)\}$. The final loss term maximizes the utility-weighted KL divergence over trajectories τ from the joint policy, $\mathcal{L}_{\text{Causal}}(\phi) = -\mathbb{E}_{\tau\sim\pi_{\theta,\phi}}[\,\mathcal{U}_t \cdot \text{ICE}_t\,]$.

The utility signal $\mathcal{U}_t$ is treated as a constant during the optimization of ϕ by detaching it from the computation graph. This guarantees that the gradient only flows through the KL divergence term, correctly isolating the effect of rewarding the IA for how its message m_t influences the CA's policy, rather than for influencing the value estimates that comprise the utility signal itself.

3.3 Multi-Environment Training for Generalization

CORAL aims to produce a communicative prior that is not overfitted to a single task but is instead broadly applicable. It is well-established that deep RL agents can achieve high performance by overfitting to the specific stochasticity of their training environments, yet fail to generalize to slightly different, unseen situations [Cobbe *et al.*, 2020].

To directly address this challenge, our pre-training regime moves beyond single-task optimization and instead trains the CORAL agents on a diverse, discrete set of tasks, $\mathcal{T} = \{\text{env}_1, \ldots, \text{env}_K\}$, that share common entities and dynamics but require different solutions. Our implementation is based on modern, vectorized training within a single process. Inspired by the architectural principles of large-scale distributed agents [Espeholt *et al.*, 2018] and using an implementation pattern similar to high-performance JAX-native frameworks like PureJaxRL [Lu *et al.*, 2022], we leverage the `vmap` transformation in JAX [Bradbury *et al.*, 2018] to run N parallel environments simultaneously on a single accelerator.

At the beginning of each rollout, each of the N parallel environment instances is randomly assigned a task from our distribution $\mathcal{T}$. This ensures that every gradient update batch contains a rich mixture of experiences from across the task family. By training on this diverse data stream, we prevent catastrophic forgetting and force the Information Agent to learn an abstract dynamics model, capturing the fundamental rules and entities common across environments rather than memorizing the specifics of any single instance. This multi-task regime is critical for learning a communicative representation that can serve as a truly generalizable prior.

3.4 Deployment for Rapid In-Context Adaptation

At deployment time, we fix the parameters ϕ of the Information Agent. IA effectively becomes a frozen deterministic module providing a continuous stream of contextual information. As such, we can precisely evaluate our main hypothesis that a trained communication protocol allows the Control Agent to adapt fast and with few samples.

We formalize this evaluation through the lens of in-context reinforcement learning [Lee *et al.*, 2023; Laskin *et al.*, 2023]. The Control Agent learning can be cast as inference of optimal policy in a new environment conditioned on the history of observations and messages constituting the context.

4 Experimental Results

We validate CORAL across three distinct regimes to assess its generalization capabilities across different data modalities. First, using the `Navix` grid-world suite [Pignatelli *et al.*, 2024], we assess **Online Adaptation**, measuring the agent's ability to learn unseen tasks from scratch, and **Zero-Shot Transfer**, evaluating robustness to increased task complexity without further training. Second, we benchmark **Offline Generalization** on the D4RL continuous control suite [Fu *et al.*, 2020]. Our analysis compares CORAL against standard PPO and monolithic World Model baselines in the online settings, while extending the comparison to state-of-the-art In-Context RL methods, including Decision Transformer (DT) and Agentic Transformer (AT), for the offline benchmarks.

Environment	Max. Perf	CORAL		PPO		PPO-TX		World Model	
		TTT	SR	TTT	SR	TTT	SR	TTT	SR
DoorKey 8x8	1.00	$\mathbf{1.0 \pm 0.2}^{\dagger}$	100%	2.3 ± 0.3	100%	3.9 ± 0.7	100%	3.0 ± 0.4	89%
Crossings S11N5	1.00	$\mathbf{1.3 \pm 0.1}^{\dagger}$	100%	1.9 ± 0.3	100%	3.9 ± 0.4	73%	6.9 ± 1.2	19%
DynObs 6x6 - Random	0.74	1.5 ± 0.2	100%	1.5 ± 0.1	97%	4.4 ± 0.2	36%	$\mathbf{1.3 \pm 0.1}$	97%
LavaGap S7	1.00	$\mathbf{0.3 \pm 0.0}^{\dagger}$	95%	0.4 ± 0.0	95%	0.6 ± 0.1	79%	0.4 ± 0.0	94%
Empty 16x16	1.00	$\mathbf{0.5 \pm 0.1}$	100%	0.7 ± 0.1	100%	2.0 ± 0.3	100%	$\mathbf{0.5 \pm 0.1}$	100%
DynObs 16x16	0.85	$\mathbf{7.0 \pm 0.6}^{\dagger}$	45%	7.9 ± 0.5	36%	8.2 ± 0.2	30%	8.5 ± 0.7	30%

Table 1: Time-to-threshold analysis (TTT, millions of steps) $\pm$ 95% confidence interval (CI) to reach 90% of maximum performance; SR: success rate. **Bold**: best per environment. †: significant improvement over next-best (Welch's t-test, $p < 0.05$).

4.1 Accelerated Control Learning via In-Context Communication

Protocol. To benchmark online in-context adaptation, we isolate source and target environments during adaptation. In this Accelerated Adaptation regime, we train an Information Agent as a single foundation model on a distribution of source tasks (*e.g.*, *navigation, lava, object-centric, dynamic-obstacle, etc.*), then freeze its weights (see Appendix C.1 for the complete pre-training set of environments). Next, we place a fresh, randomly initialized Control Agent into a distribution of unseen target tasks (*e.g.*, *DoorKey-8x8*) which learns to solve the task by consuming the frozen IA's communicative context alone. This clearly isolates the sample-efficiency boost granted by the communicative prior, as the agent must learn from zero examples. This contrasts with our Zero-Shot Generalization section, which examines fixed policies with no further adaptation.

Our first set of experiments analyzes whether a frozen pretrained CORAL IA can accelerate the learning of a randomly initialized CA from scratch on new tasks. We randomly initialize a new CA from scratch and evaluate it together with the frozen IA in previously unseen environments.

Fig. 2 demonstrates a large improvement in sample efficiency. The CORAL agent is able to achieve near-optimal rewards roughly twice as fast as the PPO agent in environments like *DoorKey 8x8* and *Crossings S11N5*. This improvement is even more dramatic when we begin to evaluate our method on harder or larger variants of the training environments. Consider Dynamic Obstacles 16x16 (*DynObs 16x16*), a much larger and more complicated environment than the environment of the same nature seen by agents during pretraining. While the baseline improves marginally over Random performance, CORAL is able to bootstrap the CA to effectively solve the task. This suggests that the learned communicative prior can provide strong yet robust guidance for the agent to avoid the exploration difficulty posed by these sparse-reward navigation tasks, even as the problem scales up.

To better quantify this improvement in sample efficiency, we conduct a time-to-threshold (TTT) analysis on each environment. We calculate the mean number of environment timesteps it takes each method to reach 90% of that agent's maximum achievable performance in each environment. We report TTT results in Table 1. As shown, CORAL is able to meet this performance threshold significantly faster than either baseline across most environments. CORAL reaches our target performance after significantly fewer timesteps in *Doorkey 8x8*, 2.3 and 3 times faster than PPO and the World Model baseline, respectively. Similarly, we see a 1.5-fold and 5+ fold improvement in TTT over PPO and the World Model baseline in *Crossings S11N5*. Lastly, we report the success rate (SR) to show CORAL's increased likelihood of actually solving these tasks. CORAL shows a 25% relative improvement in SR over PPO and 50% over WM in *DynObs16x16*, the single environment where WM performed the best relative to PPO.

These consistent results indicate that the improvement from our framework stems from the fact that it allows agents to communicate rather than using a different architecture. The gap in performance between CORAL and PPO highlights the strong learning signal provided by the message stream for agent exploration in sparse-reward environments. Moreover, CORAL significantly outperforms the architecturally equivalent World Model baseline, also highlighting the importance of our pre-training strategy. By pre-training the Information Agent with self-supervised objectives that do not directly depend on maximizing task reward, we are able to align the CA and IA onto a communication protocol that is more general than the task-specific representations learned by WM.

4.2 Integrated In-Context Communication and Control for Zero-Shot Generalization

Beyond accelerating learning from scratch, we investigate the robustness of the learned policies in a challenging zero-shot transfer setting where no further training is allowed. For this evaluation, we use the same generalist Information Agent pretrained on our full task distribution $\mathcal{T}$. We then pre-train the CORAL CA and baselines on a specific, simpler source task (e.g., *DoorKey6x6*). For the Decision Transformer (DT) baseline, we trained the model on offline trajectories collected from the converged PPO policy on these source tasks. Finally, we freeze all agent parameters and evaluate their performance on more complex, unseen target environments (e.g., *DoorKey8x8*). These experiments measure how well the learned policy generalizes to configurations of increased scale and complexity.

As shown in Table 2, we see that the CORAL architecture allows for strong zero-shot generalization performance in comparison to our baselines. The CORAL agent that was pre-trained on both *DoorKey6x6* and *LavaGapS6* obtains a much higher average return than both the PPO and World Model agents when placed in the more challenging environments of *DoorKey8x8* and *LavaGapS7*, respectively. CORAL also outperforms the DT, suggesting that the communicative prior generalizes better to dynamic shifts than purely sequence-based cloning of source task demonstrations in ICRL settings.

Environment	CORAL	PPO	PPO-TX	WM	DT
DoorKey8x8	$\mathbf{0.95 \pm 9e^{-4}}^{\dagger}$	$0.78 \pm 6e^{-4}$	$0.81 \pm 7e^{-4}$	$0.86 \pm 7e^{-3}$	$0.84 \pm 1e^{-3}$
CrossingsS11N5	$0.45 \pm 9e^{-3}$	$0.48 \pm 8e^{-3}$	$\mathbf{0.49 \pm 6e^{-3}}$	$0.20 \pm 5e^{-3}$	$0.43 \pm 8e^{-3}$
DynObs6x6-Random	$\mathbf{0.64 \pm 2e^{-3}}^{\dagger}$	$0.43 \pm 2e^{-3}$	$0.54 \pm 4e^{-3}$	$0.33 \pm 7e^{-3}$	$0.61 \pm 3e^{-3}$
LavaGapS7	$\mathbf{0.77 \pm 6e^{-3}}^{\dagger}$	$0.63 \pm 6e^{-3}$	$0.63 \pm 5e^{-3}$	$0.65 \pm 5e^{-3}$	$0.74 \pm 9e^{-3}$
Empty16x16	$\mathbf{0.90 \pm 4e^{-3}}$	$0.86 \pm 3e^{-3}$	$0.87 \pm 5e^{-3}$	$0.88 \pm 4e^{-3}$	$\mathbf{0.90 \pm 5e^{-3}}$
DynObs16x16	$\mathbf{0.52 \pm 6e^{-3}}^{\dagger}$	$0.50 \pm 5e^{-3}$	$0.47 \pm 8e^{-3}$	$0.16 \pm 6e^{-3}$	$0.43 \pm 4e^{-3}$

Table 2: Zero-shot performance with frozen weights (pre-trained on simpler source tasks). Mean episodic return ± 95% CI over 1M steps.

Dataset	Environment	TD3+BC	DT	AT	AD	CORAL
Medium-Expert	HalfCheetah	96.59	93.40	95.81 ± 0.25	94.21 ± 0.46	**96.67 ± 0.24**
Medium-Expert	Hopper	113.22	111.18	115.92 ± 1.26	108.32 ± 0.95	**116.06 ± 1.02**
Medium-Expert	Walker	112.21	108.71	**114.87 ± 0.56**	111.36 ± 0.46	114.48 ± 0.65
Medium	HalfCheetah	48.93	42.73	45.12 ± 0.34	42.28 ± 1.18	**49.01 ± 0.29**
Medium	Hopper	70.44	69.42	70.45 ± 0.45	72.58 ± 0.54	**73.15 ± 0.48**
Medium	Walker	86.91	74.70	88.71 ± 0.55	85.96 ± 0.46	**88.86 ± 0.41**
Medium-Replay	HalfCheetah	45.84	40.31	**46.86 ± 0.33**	41.28 ± 0.21	45.91 ± 0.37
Medium-Replay	Hopper	98.12	88.74	96.85 ± 0.41	91.32 ± 0.66	**98.33 ± 0.72**
Medium-Replay	Walker	91.17	68.22	**92.32 ± 1.21**	89.21 ± 1.42	86.74 ± 1.13
Total Average		84.83	77.49	85.21	82.84	**85.47**

Table 3: Normalized scores on D4RL continuous control tasks. CORAL is trained fully offline.

4.3 D4RL Results

Beyond online interaction, we benchmark CORAL against D4RL [Fu *et al.*, 2020]. As per the standard offline ICRL setup [Moeini *et al.*, 2025; Chen *et al.*, 2021], we approach it as completely offline. In this setting, the IA learns about world dynamics and reward from the offline transitions alone, creating a distilled dataset as a communicative prior. Meanwhile, the CA learns a message-conditioned policy concurrently. Table 3 displays CORAL's total average normalized score of **85.47**, beating the DT (77.49) and achieving comparable performance to TD3+BC (84.83), and the Agentic Transformer (AT) [Liu and Abbeel, 2023]. CORAL also notably outperforms other algorithms on the difficult *Medium-Replay* datasets (**98.33** on Hopper-M-Replay vs. 88.74 for DT and AT's 96.85), indicating CORAL's communicative prior is able to discard noise from sub-optimal demonstrations.

4.4 Emergent Communicative Protocol Analysis

To verify that communication is causally responsible for the observed performance gains, we study the protocol's Instantaneous Causal Effect (Eq. ICE). A high ICE value indicates that the message is causing a decisive shift in the CA's behavior.

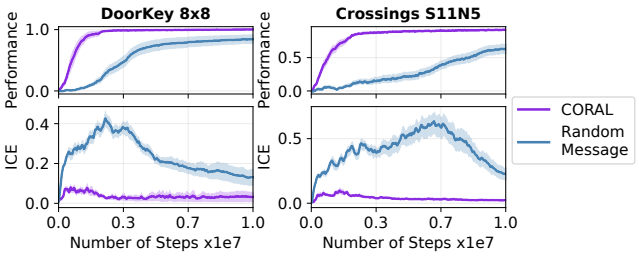

Figure 3: Communication Impact on Adaptation. Performance (top) and corresponding ICE metric (bottom) for unseen tasks.

Fig. 3 shows the mean and 95% CI of ICE averaged across all steps of CORAL and one CA trained with random messages. Although these random messages provide maximal ICE by completely distracting the agent, they provide no beneficial learning signal. Instead, CORAL's ICE increases with performance as the CA learns to trust the IA, and decreases as the learned policy is fixed and the message becomes expected information. The shape of ICE going high when learning and low when converged suggests that the CORAL protocol provides learning regularization by only asserting a strong influence when the agent is uncertain.

4.5 Ablation Study

We investigate the contribution of our specific training objectives to the emergent communication protocol (see Appendix F for full training curves).

Effect of Coherence Loss. We hypothesized that smooth transitions in the message space reduce the variance in the CA's value estimation. Our tests, in Fig. 4, confirm that $\mathcal{L}_{\text{Coh}}$ is an essential regularizer. Without it leads to high variance, slower learning, and poorer performance across all environments.

Effect of Causal Influence Loss. We also ablate the requirement for the Causal Influence objective by training a version of CORAL without incentivizing the IA to influence the CA's policy. As shown in Fig. 5, removing $\mathcal{L}_{\text{Causal}}$ hurts sample efficiency and final performance in some cases. This shows that having only a "passive" world model trained to describe the dynamics faithfully is not enough to boost downstream tasks. The causal loss incentivizes pragmatic communication where only messages that cause beneficial shifts in policy are communicated, rather than wastefully encoding easily memorizable but ineffective static environmental features.

5 Related Works

In-Context RL. In-context learning, a concept recently popularized by large language models (LLMs) research and

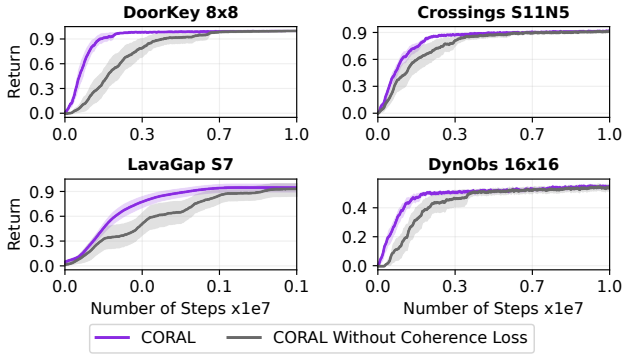

Figure 4: Ablation of $\mathcal{L}_{\text{Coh}}$. Comparison between CORAL and an ablation variant without coherence loss.

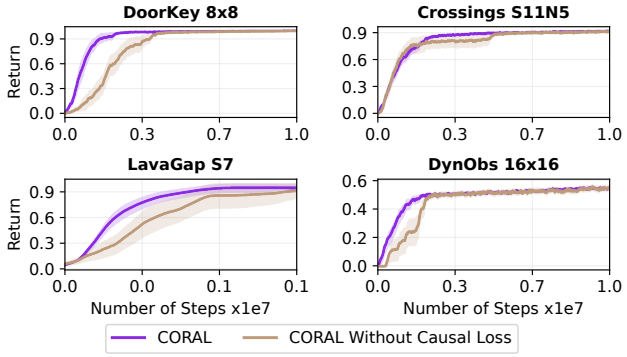

Figure 5: Ablation of $\mathcal{L}_{\text{Causal}}$. Comparison between CORAL and an ablation variant without causal loss.

applications [Dong *et al.*, 2024], refers to the ability to learn from a few examples in the context and adapt without changing model weights. The idea of few-shot adaptation in ICRL aligns with earlier work on gradient-free meta learning, where recurrent neural networks, aiming to extract task-specific hidden context from historical interactions, are incorporated into RL training [Wang *et al.*, 2016; Duan *et al.*, 2016]. Recent advancements on ICRL leverage transformers and LLMs to extract context variables (e.g., hidden states) from historical trajectories [Lee *et al.*, 2023; Krishnamurthy *et al.*, 2024]. Pre-training is key to the success in in-context adaptation. Most existing works fall within the categories of (self-)supervised pre-training and reinforcement pre-training [Moeini *et al.*, 2025].

Supervised pre-training, similar to imitation learning, encourages agents to find trajectories in the offline data that are similar to the current test task and to imitate offline actions for better generalization [Raparthy *et al.*, 2024; Xu *et al.*, 2022]. Self-supervised approaches, such as decision transformers [Chen *et al.*, 2021; Liu and Abbeel, 2023; Huang *et al.*, 2024], leverage the transformer's autoregressive generation to predict the next action based on state observations and reward feedback. Even though some advanced techniques, such as hindsight information matching [Furuta *et al.*, 2021; Li *et al.*, 2025], help improve out-of-distribution generalization, the (self-)supervised approaches largely depend on the quality of the offline data.

Our proposed CORAL belongs to the class of reinforcement pre-training, where the policy is not purely extracted from static, offline data but also from live interactions with a variety of environments. Compared with existing works on transformer-based ICRL [Bauer *et al.*, 2023; Grigsby *et al.*, 2024; Wang *et al.*, 2024], which integrate context learning into policy learning, our work further decouples the two learning processes and adopts a *hybrid* approach that combines supervised pre-training for world model training and reinforcement pre-training for control policies.

World Models. World models (WMs), typically represented by generative neural networks, serve as the agents' mental models of the world based on what they can perceive, which helps agents to conceive future consequences of their actions, thus planning in sequential decision-making [Ha and Schmidhuber, 2018]. Some early efforts focus on building visual WMs using autoencoders and recurrent neural networks [Ha and Schmidhuber, 2018; Hafner *et al.*, 2019; Hafner *et al.*, 2020; Hafner *et al.*, 2025], where WMs predict the next latent representation for planning purposes based on those of previous image inputs. In more structured tasks (e.g., Go), simple predictive models plus Monte Carlo tree search (MCTS) also yield superior performance [Schrittwieser *et al.*, 2020; Schrittwieser *et al.*, 2021; Hammar *et al.*, 2025].

Our CORAL aligns with the recent trend of transformer-based WMs [Micheli *et al.*, 2023; Robine *et al.*, 2023] for its improved sample efficiency, compared with RNN-based ones. However, these transformer WMs, like most WMs, directly incorporate the general-purpose dynamics modeling learning with task-specific RL, where agents treat the transformer-generated latent representations as the new state variables. In contrast, our approach decouples the two learning tasks, and the transformer is only tasked with understanding the shared environment dynamics and communicating them to the control agent as side information or context. Most relevant to ours is [Toledo and Prorok, 2024], which studies communication-based decentralized WMs. This work proposes to equip each agent with a WM in a multi-agent RL task, where WMs first communicate with each other on future predictions. CORAL, instead, focuses on tackling a single-agent RL from an emergent communication perspective, where the transformer WM is still the mental model without actually optimizing the policy.

6 Conclusion

This work introduces CORAL, a framework for learning a communicative prior for in-context reinforcement learning. By pre-training an Information Agent as a communicative world model with objectives decoupled from direct task rewards, we learn a reusable protocol that effectively conveys relevant knowledge to instruct a control agent's actions. Our experiments show that using this pre-trained communicative prior enables improvements in sample efficiency for several online adaptation and better zero-shot generalization in unseen, sparse-reward tasks. Additionally, we show CORAL achieves competitive results on offline RL tasks, suggesting emergent communication is promising not just for coordination but as a general method for rapid adaptation.

Acknowledgments

Juntao Chen acknowledges support from the Fordham AI Research (FAIR) Grant and the Fordham Faculty Research Grant.

References

[Bauer *et al.*, 2023] Jakob Bauer, Kate Baumli, et al. Human-timescale adaptation in an open-ended task space. In *ICML*, 2023.

[Bradbury *et al.*, 2018] James Bradbury, Roy Frostig, Peter Hawkins, et al. JAX: composable transformations of Python+NumPy programs. http://github.com/jax-ml/jax, 2018.

[Chen *et al.*, 2021] Lili Chen, Kevin Lu, Aravind Rajeswaran, et al. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, volume 34, pages 15084–15097, 2021.

[Chen *et al.*, 2024] Jiayu Chen, Bhargav Ganguly, Yang Xu, et al. Deep generative models for offline policy learning: Tutorial, survey, and perspectives on future directions. *TMLR*, 2024.

[Cobbe *et al.*, 2020] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *ICML*, pages 2048–2056. PMLR, 2020.

[Crawford and Sobel, 1982] Vincent P Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431, 1982.

[Dong *et al.*, 2024] Qingxiu Dong, Lei Li, Damai Dai, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Duan *et al.*, 2016] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RLˆ 2: Fast reinforcement learning via slow reinforcement learning. *arXiv:1611.02779*, 2016.

[Espeholt *et al.*, 2018] Lasse Espeholt, Hubert Soyer, Remi Munos, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *ICML*, pages 1407–1416, 2018.

[Eysenbach *et al.*, 2022] Benjamin Eysenbach, Alexander Khazatsky, Sergey Levine, and Ruslan Salakhutdinov. Joint model-policy optimization of a lower bound for model-based RL. In *NeurIPS*, 2022.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, volume 70, pages 1126–1135, 2017.

[Fu *et al.*, 2020] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[Furuta *et al.*, 2021] Hiroki Furuta, Yutaka Matsuo, and Shixiang Shane Gu. Generalized Decision Transformer for Offline Hindsight Information Matching. In *ICLR*, 2021.

[Grigsby *et al.*, 2024] Jake Grigsby, Linxi Fan, and Yuke Zhu. AMAGO: Scalable in-context rl for adaptive agents. In *ICLR*, 2024.

[Ha and Schmidhuber, 2018] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, volume 31, 2018.

[Hafner *et al.*, 2019] Danijar Hafner, Timothy Lillicrap, Ian Fischer, et al. Learning latent dynamics for planning from pixels. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565. PMLR, 09–15 Jun 2019.

[Hafner *et al.*, 2020] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

[Hafner *et al.*, 2025] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

[Hammar *et al.*, 2025] Kim Hammar, Tao Li, Rolf Stadler, and Quanyan Zhu. Adaptive security response strategies through conjectural online learning. *IEEE Transactions on Information Forensics and Security*, 20:4055–4070, 2025.

[Huang *et al.*, 2024] Sili Huang, Jifeng Hu, Zhejian Yang, et al. Decision mamba: Reinforcement learning via hybrid selective sequence modeling. In *NeurIPS*, 2024.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. pmlr, 2015.

[Kaelbling *et al.*, 1998] Leslie Kaelbling, Michael Littman, and Anthony Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101:99–134, 1998.

[Krishnamurthy *et al.*, 2024] Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can LLM explore in-context? In *ICML 2024 Workshop on In-Context Learning*, 2024.

[Laskin *et al.*, 2023] Michael Laskin, Luyu Wang, Junhyuk Oh, et al. In-context reinforcement learning with algorithm distillation. In *ICLR*, 2023.

[Lee *et al.*, 2023] Jonathan Lee, Annie Xie, Aldo Pacchiano, et al. Supervised pretraining can learn in-context reinforcement learning. In *NeurIPS*, 2023.

[Li and Zhu, 2023] Tao Li and Quanyan Zhu. On the price of transparency: A comparison between overt persuasion and covert signaling. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4267–4272, 2023.

[Li *et al.*, 2023] Tao Li, Haozhe Lei, and Quanyan Zhu. Self-adaptive driving in nonstationary environments through conjectural online lookahead adaptation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7205–7211, 2023.

[Li *et al.*, 2024] Tao Li, Henger Li, Yunian Pan, Tianyi Xu, Zizhan Zheng, and Quanyan Zhu. Meta Stackelberg game: Robust federated learning against adaptive and mixed poisoning attacks. *arXiv:2410.17431*, 2024.

[Li *et al.*, 2025] Tao Li, Juan Guevara, Xinhong Xie, and Quanyan Zhu. Self-confirming transformer for belief-conditioned adaptation in offline multi-agent reinforcement learning. In *Proceedings of the Seventh Workshop on Adaptive and Learning Agents*, pages 1–10, 2025.

[Liu and Abbeel, 2023] Hao Liu and Pieter Abbeel. Emergent agentic transformer from chain of hindsight experience. In *International Conference on Machine Learning*, pages 21362–21374. PMLR, 2023.

[Lowe *et al.*, 2019] Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 693–701, 2019.

[Lu *et al.*, 2022] Chris Lu, Jakub Kuba, Alistair Letcher, et al. Discovered policy optimisation. *NeurIPS*, 35:16455–16468, 2022.

[Micheli *et al.*, 2023] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *ICLR*, 2023.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Moeini *et al.*, 2025] Amir Moeini, Jiuqi Wang, Jacob Beck, et al. A survey of in-context reinforcement learning. *arXiv:2502.07978*, 2025.

[Pan *et al.*, 2025] Yunian Pan, Tao Li, and Quanyan Zhu. Model-agnostic hessian-free meta-policy optimization via zeroth-order estimation: A linear quadratic regulator perspective. *Dynamic Games and Applications*, pages 1–40, 2025.

[Pignatelli *et al.*, 2024] Eduardo Pignatelli, Jarek Liesen, Robert Tjarko Lange, et al. Navix: Scaling minigrid environments with jax. *arXiv:2407.19396*, 2024.

[Raparthy *et al.*, 2024] Sharath Chandra Raparthy, Eric Hambro, Robert Kirk, Mikael Henaff, and Roberta Raileanu. Generalization to new sequential decision making tasks with in-context learning. In *ICML*, ICML, pages 42138–42158, 2024.

[Reed, 2022] Scott Reed. A generalist agent. *Transactions on Machine Learning Research*, 11 2022.

[Robine *et al.*, 2023] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *The Eleventh International Conference on Learning Representations*, 2023.

[Schrittwieser *et al.*, 2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[Schrittwieser *et al.*, 2021] Julian Schrittwieser, Thomas Hubert, Amol Mandhane, et al. Online and offline reinforcement learning by planning with a learned model. In *NeurIPS*, volume 34, pages 27580–27591, 2021.

[Schulman *et al.*, 2015] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv:1506.02438*, 2015.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[Schwarzer *et al.*, 2021] Max Schwarzer, Nitarshan Rajkumar, Michael Noukhovitch, et al. Pretraining representations for data-efficient reinforcement learning. *NeurIPS*, 34:12686–12699, 2021.

[Stooke *et al.*, 2021] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *ICML*, volume 139 of *PMLR*, pages 9870–9879, 2021.

[Sukhbaatar *et al.*, 2016] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *NeurIPS*, Red Hook, NY, USA, 2016. Curran Associates Inc.

[Toledo and Prorok, 2024] Edan Toledo and Amanda Prorok. Codreamer: Communication-based decentralised world models. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *NeurIPS*, 30, 2017.

[Wang *et al.*, 2016] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, et al. Learning to reinforcement learn. *arXiv:1611.05763*, 2016.

[Wang *et al.*, 2024] Jiuqi Wang, Ethan H Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers learn temporal difference methods for in-context reinforcement learning. In *ICML 2024 Workshop on In-Context Learning*, 2024.

[Xu *et al.*, 2022] Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, and Chuang Gan. Prompting decision transformer for few-shot policy generalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24631–24645. PMLR, 17–23 Jul 2022.

[Zhu *et al.*, 2024] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4, 2024.