In-Context Reinforcement Learning via Communicative World Models

Fernando Martinez-Lopez[1*], **Tao Li**[2†], **Yingdong Lu**[3], **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
[1]{fmartinezlopez, jchen504}@fordham.edu, [2]li.tao@cityu.edu.hk, [3]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 decoupling 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 to maximize task reward, but to build a world model and distill 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 entirely unseen sparse-reward 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). A general-purpose RL algorithm spares one from hand-crafting domain-specific inductive biases in training (Hessel et al. 2019; Li et al. 2025b) and enables fast adaptation to unseen and complex environments (Li et al. 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. 2025a), which reveals task-related information that improves generalization (Li, Lei, and Zhu 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, Abbeel, and Levine 2017; Fallah et al. 2021; Li et al. 2024b; Pan, Li, and Zhu 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, 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 contextualizersthan 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 decouples the representation and policy learning and formulates ICRL as a two-agent emergent communication problem (Lowe et al. 2019; Zhu, Dastani, and

[*]This research was supported by the AI Summer Research Fellowship from Fordham University's GSAS

[†]Corresponding author

[1]Code and experimental details are available at `https://github.com/fernando-ml/CORAL`.

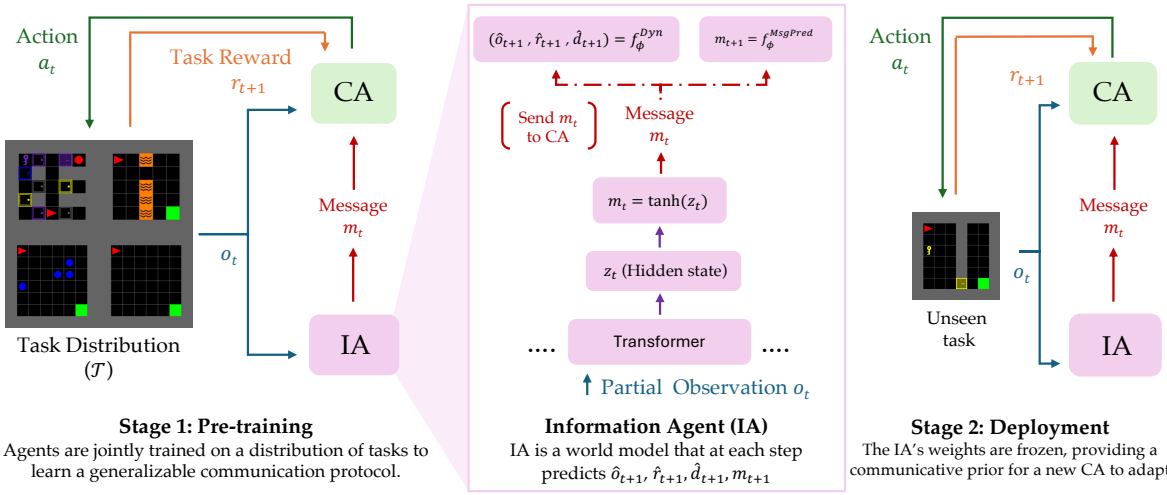

Figure 1: **Overview of the CORAL framework for learning and deploying a communicative prior. (Left) Pre-training.** An Information Agent (IA, parameterized by ϕ) and a Control Agent (CA, parameterized by θ) are trained jointly on a distribution of tasks $\mathcal{T}$. The IA, a Transformer-based model detailed in the center panel, processes a partial observation o_t to produce a message m_t. The CA is a standard PPO agent that conditions its policy on both o_t and m_t to select an action a_t, learning from the task reward r_{t+1}. The IA is trained with a set of self-supervised and communication-based objectives. The pre-trained IA is frozen and paired with a new, randomly initialized CA. **(Right) Deployment**. The IA provides a continuous stream of messages, acting as a communicative prior to accelerate the CA's adaptation to an unseen task.

Wang 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, where CORAL yields improved sample-efficiency (5X faster than vanilla WM) and facilitates zero-shot generalization.

Preliminaries

We formulate RL tasks as Partially Observable Markov Decision Processes (POMDP) (Kaelbling, Littman, and Cassandra 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 POMDP under cheap-talk setting is mathematically a communication game (Crawford and Sobel 1982; Lowe et al. 2019; Li and Zhu 2023; Yang, Li, and Zhu 2025). 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, Szlam, and Fergus 2016) or a totally different one (Crawford and Sobel 1982; Li and Zhu 2023).

Methodology

The challenge of enabling intelligent agents to generalize and rapidly adapt to new situations is a central pursuit in reinforcement learning. Standard end-to-end paradigms, where a single reward signal shapes a monolithic network, can be remarkably effective (Mnih et al. 2015) but often yield representations that are overly specialized to the training task, hindering generalization. This has motivated research into decoupled approaches where general-purpose representations are learned via self-supervised objectives (Stooke et al. 2021; Schwarzer et al. 2021). While these methods can produce

robust features, they may not be optimally tailored for the specific downstream control task.

Overview We propose CORAL (Communicative Representation for Adaptive RL), a framework that navigates a principled middle ground. CORAL conceptualizes the learning problem through a functional separation of representation and control. A dedicated Information Agent (IA) learns a dynamics model, while a separate Control Agent (CA) is tasked with maximizing rewards. This structure allows the IA's learned protocol to serve as a powerful contextual prior for the CA, enabling rapid in-context adaptation.

CORAL's design is centered on the asymmetric learning objectives of the two agents. The CA is a standard RL agent driven by extrinsic task reward, whereas the IA's learning objective is deliberately decoupled from this reward signal. The IA is responsible for building a predictive model of the environment's dynamics and distilling this understanding into a communicative representation. For this decoupling to be effective, the IA's communication must be explicitly shaped to benefit the CA's control policy.

This is achieved through a composite loss that combines self-supervised dynamics modeling with a causal influence objective. The IA is trained to predict future observations, rewards, and termination signals, which grounds its internal representations in the physical reality of the environment. Simultaneously, its Causal Influence Loss shapes the content of its messages. This objective encourages the IA to generate communications that produce a large, decisive shift in the CA's policy, but only when that shift is correlated with high-utility outcomes. The utility is measured by a hybrid signal combining the long-term Generalized Advantage Estimate (GAE) with the immediate change in the CA's own value estimate, effectively teaching the IA to communicate information that is both immediately and prospectively useful. This principled separation allows the IA to form a general representation, which serves as a powerful, pre-trained communicative prior during deployment.

CORAL Agents and Emergent Communication

CORAL is realized through two distinct neural network architectures that interact at each timestep t. Both agents receive the same partial observations $o_t \in \mathbb{R}^{D_{\text{obs}}}$, but serve specialized roles. The Information Agent (IA), parameterized by ϕ, is designed to integrate historical information and model environmental dynamics. It is implemented as a Transformer-based architecture (Vaswani et al. 2017) that operates on a context window of the most recent L observation embeddings, denoted as $c_t = (e_{t-L+1}, \ldots, e_t)$. At each timestep, the observation o_t is transformed into an embedding e_t through a linear layer, updating the context window via a sliding buffer. The sequence c_t is then processed through a series of self-attention and feed-forward layers with residual connections, allowing every token in the context to attend to every other. This produces a sequence of contextualized output vectors.

We extract the final output vector, z_t, which corresponds to the most recent observation, as the IA's state representation. This architecture allows the IA to reason about temporal dependencies within its observation history when constructing its communication. The message is then generated by a final linear layer with a $\tanh$ activation: $m_t = \tanh(f_\phi^{\text{Msg}}(z_t))$. The $\tanh$ function bounds the message vector, stabilizing the communication channel. Crucially, the IA's learning is not directly driven by extrinsic task rewards but by a set of self-supervised and communication-centric objectives.

The Control Agent (CA), parameterized by θ, is responsible for action and control. It is a standard actor-critic agent trained with Proximal Policy Optimization (PPO) (Schulman et al. 2017). Its policy $\pi_\theta(\cdot|o_t, m_t)$ and value function $V_\theta(o_t, m_t)$ are conditioned on both its observation and the message. Its sole objective is to maximize the task reward.

Pre-training the Communicative Representation

The parameters ϕ and θ of both agents are optimized jointly in pre-training. The learning objectives are designed to produce an IA that understands the world's dynamics and can communicate that understanding effectively, and a CA that can leverage this 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 grounds the IA's representations by training it to predict the consequences of the CA's actions. Using the message m_t and the action a_t taken by the CA, the IA predicts the next observation $\hat{o}_{t+1}$, reward $\hat{r}_{t+1}$, and termination probability $\hat{d}_{t+1}$; see Appendix C for details.

$$\mathcal{L}_{\text{Dyn}} = \mathbb{E}_t[\underbrace{\|\hat{o}_{t+1} - o_{t+1}\|^2}_{①} + \underbrace{(\hat{r}_{t+1} - r_{t+1})^2}_{②} + \underbrace{\text{BCE}(\hat{d}_{t+1}, d_{t+1})}_{③}],$$

where predictions are generated by prediction heads (f_ϕ^{Dyn}) conditioned on the message m_t and action a_t: $(\hat{o}'_{t+1}, \hat{r}_{t+1}, \hat{d}_{t+1}) = f_\phi^{\text{Dyn}}(m_t, a_t)$. ① and ② use mean-square error since observation and action predictions are continuous, whereas ③ considers binary cross-entropy (BCE) loss for probabilistic outputs.

Temporal Coherence Loss ($\mathcal{L}_{\text{Coh}}$): To promote temporal coherence in communication, the IA is also trained to predict the next 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:

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

Here, $\pi_\theta(\cdot|o_t, \mathbf{0})$ is 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 GAE (Schulman et al. 2015), A_t, which provides a low-variance estimate of the long-term, empirical advantage 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 $\mathcal{U}_t$:

$$\mathcal{U}_t = \max\{0, \alpha \cdot \text{norm}(\Delta V_t) + (1 - \alpha) \cdot \text{norm}(A_t)\}.$$

The final loss term maximizes the utility-weighted KL divergence over trajectories τ from the joint policy,

$$\mathcal{L}_{\text{Causal}}(\phi) = -\mathbb{E}_{\tau\sim\pi_{\theta,\phi}}\left[\,\mathcal{U}_t \cdot \text{ICE}_t\,\right].$$

The utility signal $\mathcal{U}_t$ is treated as a constant during the optimization of ϕ by detaching it from the computation graph. This ensures that the gradient only flows through the KL divergence term, correctly isolating the objective to 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.

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.

Deployment for Rapid In-Context Adaptation

During deployment, the parameters ϕ of the Information Agent are frozen. The IA now functions as a fixed, deterministic module that provides a continuous stream of contextual information. This setup allows us to rigorously test our central hypothesis that the learned communication protocol enables rapid, sample-efficient adaptation for a Control Agent.

We formalize this evaluation through the lens of in-context reinforcement learning (Lee et al. 2023; Laskin et al. 2023). The Control Agent's learning process can be viewed as an inference problem, where it must deduce the optimal policy for a new environment by conditioning on the history of observations and messages that form its context.

Experimental Results

We conduct a series of experiments designed to validate our central hypothesis: a pre-trained communicative prior, learned via CORAL, can enable rapid in-context adaptation by providing a rich, task-relevant learning signal in environments where such information is otherwise unavailable. Specifically, we demonstrate that CORAL excels in settings characterized by sparse rewards, partial observability, and long-range dependencies (challenges that are known to impede standard model-free agents). Accordingly, our evaluation is structured to first demonstrate our framework's sample efficiency compared to baselines, then to rigorously test the generalization of its learned protocol to unseen tasks, and finally to analyze the communication itself to verify it is causally responsible for the observed gains.

To this end, we perform experiments on a suite of partially observable grid-world environments using Navix (Pignatelli et al. 2024), a high-performance JAX-native implementation of MiniGrid (Chevalier-Boisvert et al. 2023). These environments are well-suited for our work as they feature sparse reward and long decision-making horizons challenges. Our pre-training distribution $\mathcal{T}$ is curated to be diverse, including tasks such as Empty, Crossings, DoorKey, LavaGap, DynamicObstacles, and randomized variants (details in Appendix C). We evaluate generalization ability by deploying agents in larger or more complex unseen versions of these tasks.

To ensure a fair and comprehensive comparison, we evaluate the performance of our CORAL agent against two strong baselines. The first is a standard PPO agent with a feed-forward policy, representing a tabula rasa learning agent,

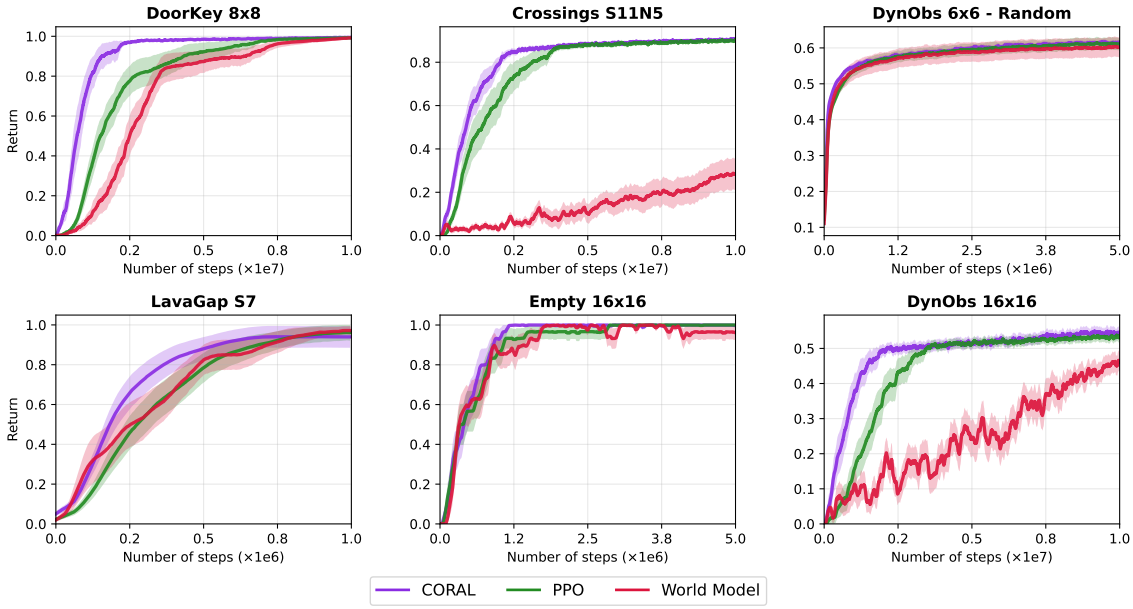

Figure 2: **In-context adaptation with a pre-trained Information Agent.** Learning curves show the mean episodic return ($\pm 95\%$ confidence interval) across 30 multiple seeds for a randomly initialized Control Agent paired with a pre-trained, frozen CORAL IA. CORAL demonstrates higher sample efficiency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

with the same architecture of CORAL's control agent utilized for these experiments. The second is a more powerful baseline we term *World Model* (WM), is a Transformer-based agent whose architecture is identical to our IA. The WM is trained with a composite objective that combines the standard PPO loss for policy and value functions with the same self-supervised self-supervised dynamics-prediction objectives ($\mathcal{L}_{\text{Dyn}}$) used by our IA.

Accelerated Control Learning via In-Context Communication

Our first set of experiments assesses the ability of the pre-trained CORAL Information Agent to accelerate the learning of a new Control Agent from scratch in unseen tasks. In this setting, the IA's parameters are frozen, and it is paired with a randomly initialized CA in a series of unseen evaluation environments. The results presented in Fig. 2, show that the CORAL-guided agent consistently and substantially outperforms both the PPO and world model baseline.

The learning curves illustrate a significant gain in sample efficiency. In tasks such as *DoorKey 8x8* and *Crossings S11N5*, the CORAL agent reaches near optimal performance in approximately half the time required by the PPO agent. This advantage is particularly pronounced when applied to more complex or larger versions of the training tasks. For instance, in Dynamic Obstacles 16x16 (*DynObs 16x16*), which is a substantially larger and more challenging variant of an environment seen during pretraining, the baselines struggle to make meaningful progress while CORAL makes the CA quickly discover a good policy. This shows that the communicative prior provides effective guidance that is not brittle, allowing a new agent to overcome the exploration challenges inherent in these sparse-reward tasks even as the scale and difficulty of the problem increase.

To quantify this sample efficiency, we perform a time-to-threshold (TTT) analysis, measuring the mean number of timesteps required for each method to reach 90% of the maximum performance achieved in each environment. As shown in Table 1, CORAL achieves the target performance threshold significantly faster than the baselines in most tasks. For example, in *Doorkey 8x8*, CORAL reaches the target performance in significantly fewer timesteps, achieving the threshold 2.3 times faster than PPO and 3 times faster than the World Model baseline. Similarly, CORAL demonstrates a 1.5-fold speedup over PPO and more than a 5-fold improvement over the World Model baseline in *Crossings S11N5*. The success rate (SR) further highlights CORAL's reliability to solve these tasks; for instance, in the most challenging environment, *DynObs16x16*, CORAL achieves a 25% relative improvement in SR over PPO and a 50% over the WM.

The consistent and significant results indicate that the benefit of our framework stem not from the use of a different architecture, but from the communication itself. The performance gap over the PPO baseline demonstrates that the message stream provides a powerful learning signal for exploration in sparse-reward settings. Furthermore, CORAL's superiority over the architecturally-equivalent World Model baseline highlights the benefit of our pre-training strategy. By training the Information Agent with objectives that are not directly tied to maximizing task reward, CORAL develops a more general communication protocol that serves as a more effective in-context learning signal than the task-specific representations learned by the end-to-end model.

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

Table 1: Time-to-threshold analysis showing mean timesteps (in millions) ± 95% confidence interval to reach 90% maximum performance. SR (%) is the success rate of runs reaching the threshold. **Bold** indicate the best method per environment. † indicates statistically significant improvement over the next-best method (Welch's t-test, $p < 0.05$).

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 learning is allowed. For this evaluation, we use the same generalist Information Agent pre-trained on our full task distribution $\mathcal{T}$. We then pre-train the CORAL CA and both baselines on specific, simpler source task (e.g., *DoorKey6x6*). Finally, we freeze all agent parameters and evaluate their performance directly on more complex, unseen target environments (e.g., *DoorKey8x8*). This set of experiments measure how well the learned policy generalize to configurations of increased scale and complexity.

The results, summarized in Table 2, demonstrate that the CORAL system exhibits superior zero-shot generalization capabilities compared to the baselines. In environments like *DoorKey8x8* and *LavaGapS7*, the pre-trained CORAL agent achieves significantly higher average returns than both the PPO and World Model agents, which were pre-trained on the same source tasks.

Environment	CORAL	PPO	WM
DoorKey8x8	$\mathbf{0.95 \pm 9e^{-4}}^{\dagger}$	$0.78 \pm 6e^{-4}$	$0.86 \pm 7e^{-3}$
CrossingsS11N5	$0.45 \pm 9e^{-3}$	$\mathbf{0.48 \pm 8e^{-3}}^{\dagger}$	$0.20 \pm 5e^{-3}$
DynObs6x6-Rand	$\mathbf{0.64 \pm 2e^{-3}}^{\dagger}$	$0.43 \pm 2e^{-3}$	$0.33 \pm 7e^{-3}$
LavaGapS7	$\mathbf{0.77 \pm 6e^{-3}}^{\dagger}$	$0.63 \pm 6e^{-3}$	$0.65 \pm 5e^{-3}$
Empty16x16	$\mathbf{0.90 \pm 4e^{-3}}^{\dagger}$	$0.86 \pm 3e^{-3}$	$0.88 \pm 4e^{-3}$
DynObs16x16	$\mathbf{0.52 \pm 6e^{-3}}^{\dagger}$	$0.50 \pm 5e^{-3}$	$0.16 \pm 6e^{-3}$

Table 2: Zero-shot performance. All agents were pre-trained on simpler source tasks and evaluated with frozen weights. Values are mean episodic return ± 95% CI over 1M steps. **Bold** indicate the best method per environment. † indicates statistically significant improvement over the next-best method (Welch's t-test, $p < 0.05$).

Analysis of the Emergent Communicative Protocol

To verify that communication is causally responsible for the observed performance gains, we analyze 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.

Fig. 3 presents the mean and 95% confidence interval of ICE for CORAL alongside a Control Agent receiving random messages. While the random messages induce a high ICE by acting as a noisy distractor, they result in no meaningful learning. In contrast, CORAL's ICE rises in lockstep with performance as the CA learns to rely on the IA's guidance, and then recedes as the policy is mastered and the message becomes confirmatory. This dynamic, where influence is high during learning and low at convergence, suggests that the CORAL protocol acts as an effective learning catalyst, providing strong guidance only when the agent is uncertain.

Related Works

In-Context Reinforcement Learning In-context learning, a concept recently popularized by large language models (LLMs) research and applications (Dong et al. 2024; Li and Zhu 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 coincides with some earlier works 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).

Thanks to the transformer architecture's sequence modeling capability (Vaswani et al. 2017), most recent advancements on ICRL leverage transformer and LLMs to extract context variables (e.g., hidden states) from historical trajectories (Lee et al. 2023; Li et al. 2025a; Krishnamurthy et al. 2024). Pre-training in ICRL is the key to its successful in-context adaptation in deployment. Most existing works fall within the categories of (self-) supervised pre-training and reinforcement pre-training (Moeini et al. 2025).

Supervised pre-training, bearing a similar spirit to imitation learning, encourages agents to find trajectories similar to the current testing task from the offline data and imitate offline actions for better generalization (Raparthy et al. 2024; Xu et al. 2022). Self-supervised approach, such as decision transformers (Chen et al. 2021; Huang et al. 2024), utilizes the transformer's auto-regressive generation and predicts the next action based on the predictions on the state observation and reward feedback. Even though some advanced techniques, such as hindsight information matching (Furuta, Matsuo, and Gu 2021; Li et al. 2025a), help improve out-of-distribution generalization, the (self-) supervised approaches largely depend on the quality of offline pertaining data.

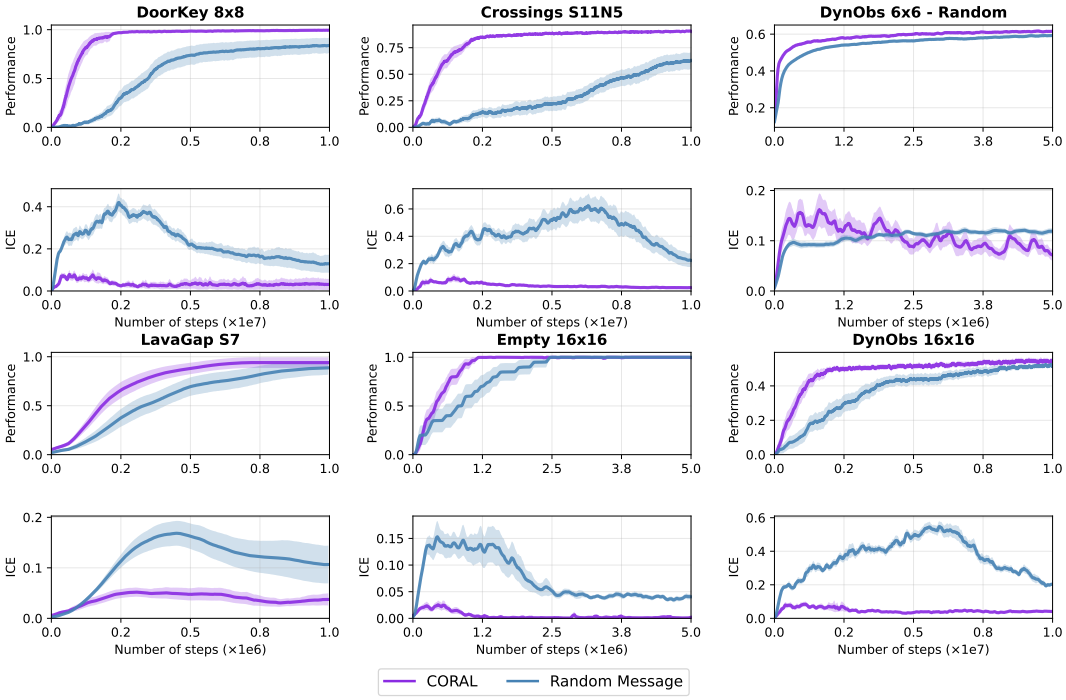

Figure 3: **Analysis of Influence of Communication during In-Context Adaptation.** Learning curves (top row) are paired with the corresponding Instantaneous Causal Effect metric (bottom row) for six unseen environments. ICE measures the per-step policy shift induced by communication (Eq. ICE). The Random Message baseline induces a high but unproductive ICE by distracting the agent. Conversely, the ICE of the CORAL agent rises as the agent learns and then recedes as the policy converges.

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, Fan, and Zhu 2024; Wang et al. 2024), which integrates the learning of context into policy learning, our work further decouples the two learning processes and adopts a *hybrid* of 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 are able to 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 (RNNs) (Ha and Schmidhuber 2018; Hafner et al. 2019, 2020, 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, 2021; Li et al. 2024a; Hammar et al. 2025).

Our CORAL aligns with the recent trend of transformer-based WMs (Micheli, Alonso, and Fleuret 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.

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 produce a generalizable protocol that effectively guides a new agent's learning process. Our experiments demonstrate that using this pre-trained communicative prior leads to significant gains in sample efficiency and zero-shot performance for new newly initialized agents and in unseen, sparse-reward tasks. This work validates the promise of using emergent communication not merely for coordination, but

as a powerful mechanism for rapid adaptation in multi-agent systems.

References

Bauer, J.; Baumli, K.; Behbahani, F.; Bhoopchand, A.; Bradley-Schmieg, N.; Chang, M.; Clay, N.; Collister, A.; Dasagi, V.; Gonzalez, L.; Gregor, K.; Hughes, E.; Kashem, S.; Loks-Thompson, M.; Openshaw, H.; Parker-Holder, J.; Pathak, S.; Perez-Nieves, N.; Rakicevic, N.; Rocktäschel, T.; Schroecker, Y.; Singh, S.; Sygnowski, J.; Tuyls, K.; York, S.; Zacherl, A.; and Zhang, L. 2023. Human-timescale adaptation in an open-ended task space. In *Proceedings of the 40th International Conference on Machine Learning*.

Bradbury, J.; Frostig, R.; Hawkins, P.; Johnson, M. J.; Leary, C.; Maclaurin, D.; Necula, G.; Paszke, A.; VanderPlas, J.; Wanderman-Milne, S.; and Zhang, Q. 2018. JAX: composable transformations of Python+NumPy programs. `http://github.com/jax-ml/jax`.

Chen, J.; Ganguly, B.; Xu, Y.; Mei, Y.; Lan, T.; and Aggarwal, V. 2024. Deep Generative Models for Offline Policy Learning: Tutorial, Survey, and Perspectives on Future Directions. *Transactions on Machine Learning Research*. Survey Certification.

Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 15084–15097.

Chevalier-Boisvert, M.; Dai, B.; Towers, M.; Perez-Vicente, R.; Willems, L.; Lahlou, S.; Pal, S.; Castro, P. S.; and Terry, J. 2023. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36: 73383–73394.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cobbe, K.; Hesse, C.; Hilton, J.; and Schulman, J. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, 2048–2056. PMLR.

Crawford, V. P.; and Sobel, J. 1982. Strategic Information Transmission. *Econometrica*, 50(6): 1431.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A Survey on In-context Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. Miami, Florida, USA: Association for Computational Linguistics.

Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. RL$^ 2$: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv*.

Espeholt, L.; Soyer, H.; Munos, R.; Simonyan, K.; Mnih, V.; Ward, T.; Doron, Y.; Firoiu, V.; Harley, T.; Dunning, I.; et al. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, 1407–1416. PMLR.

Eysenbach, B.; Khazatsky, A.; Levine, S.; and Salakhutdinov, R. 2022. Joint Model-Policy Optimization of a Lower Bound for Model-Based RL. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Fallah, A.; Georgiev, K.; Mokhtari, A.; and Ozdaglar, A. 2021. On the Convergence Theory of Debiased Model-Agnostic Meta-Reinforcement Learning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 3096–3107.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 1126–1135.

Furuta, H.; Matsuo, Y.; and Gu, S. S. 2021. Generalized Decision Transformer for Offline Hindsight Information Matching. In *International Conference on Learning Representations*.

Grigsby, J.; Fan, L.; and Zhu, Y. 2024. AMAGO: Scalable In-Context Reinforcement Learning for Adaptive Agents. In *The Twelfth International Conference on Learning Representations*.

Ha, D.; and Schmidhuber, J. 2018. Recurrent World Models Facilitate Policy Evolution. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *International Conference on Learning Representations*.

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019. Learning Latent Dynamics for Planning from Pixels. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2555–2565. PMLR.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2025. Mastering diverse control tasks through world models. *Nature*, 640(8059): 647–653.

Hammar, K.; Li, T.; Stadler, R.; Zhu, Q.; and Hammar, K. 2025. Adaptive Security Response Strategies Through Conjectural Online Learning. *IEEE Transactions on Information Forensics and Security*, 20: 4055–4070.

Hessel, M.; Hasselt, H. v.; Modayil, J.; and Silver, D. 2019. On Inductive Biases in Deep Reinforcement Learning. *arXiv*.

Huang, S.; Hu, J.; Yang, Z.; Yang, L.; Luo, T.; Chen, H.; Sun, L.; and Yang, B. 2024. Decision Mamba: Reinforcement Learning via Hybrid Selective Sequence Modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate

shift. In *International conference on machine learning*, 448–456. pmlr.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.

Krishnamurthy, A.; Harris, K.; Foster, D. J.; Zhang, C.; and Slivkins, A. 2024. Can large language models explore in-context? In *ICML 2024 Workshop on In-Context Learning*.

Laskin, M.; Wang, L.; Oh, J.; Parisotto, E.; Spencer, S.; Steigerwald, R.; Strouse, D.; Hansen, S. S.; Filos, A.; Brooks, E.; maxime gazeau; Sahni, H.; Singh, S.; and Mnih, V. 2023. In-context Reinforcement Learning with Algorithm Distillation. In *The Eleventh International Conference on Learning Representations*.

Lee, J.; Xie, A.; Pacchiano, A.; Chandak, Y.; Finn, C.; Nachum, O.; and Brunskill, E. 2023. Supervised Pretraining Can Learn In-Context Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Li, T.; Guevara, J.; Xie, X.; and Zhu, Q. 2025a. Self-Confirming Transformer for Belief-Conditioned Adaptation in Offline Multi-Agent Reinforcement Learning. In *Proceedings of the Seventh Workshop on Adaptive and Learning Agents, the Twenty Fourth International Conference on Autonomous Agents and Multiagent Systems*, 1–10. [Online] Avaiable at `https://openreview.net/forum?id=kMaYSsEwCT`.

Li, T.; Hammar, K.; Stadler, R.; and Zhu, Q. 2024a. Conjectural Online Learning with First-order Beliefs in Asymmetric Information Stochastic Games. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, IEEE CDC, 6780–6785.

Li, T.; Lei, H.; Yin, M.; and Hu, Y. 2025b. Reinforcement Learning with Physics-Informed Symbolic Program Priors for Zero-Shot Wireless Indoor Navigation. In *Reinforcement Learning Conference 2025, Inductive Biases in Reinforcement Learning Workshop*. [Online] Available at `https://arxiv.org/pdf/2506.22365`.

Li, T.; Lei, H.; and Zhu, Q. 2023. Self-Adaptive Driving in Nonstationary Environments through Conjectural Online Lookahead Adaptation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 7205–7211.

Li, T.; Li, H.; Pan, Y.; Xu, T.; Zheng, Z.; and Zhu, Q. 2024b. Meta stackelberg game: Robust federated learning against adaptive and mixed poisoning attacks. *arXiv preprint arXiv:2410.17431*. [Online] Available at `https://arxiv.org/pdf/2410.17431`.

Li, T.; Peng, G.; Zhu, Q.; and Baar, T. 2022. The Confluence of Networks, Games, and Learning a Game-Theoretic Framework for Multiagent Decision Making Over Networks. *IEEE Control Systems*, 42(4): 35–67.

Li, T.; and Zhu, Q. 2023. On the Price of Transparency: A Comparison Between Overt Persuasion and Covert Signaling. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 4267–4272.

Li, T.; and Zhu, Q. 2024. Symbiotic Game and Foundation Models for Cyber Deception Operations in Strategic Cyber Warfare. *arXiv preprint arXiv:2403.10570*. [Online] Available at `https://arxiv.org/pdf/2403.10570`.

Lowe, R.; Foerster, J.; Boureau, Y.-L.; Pineau, J.; and Dauphin, Y. 2019. On the Pitfalls of Measuring Emergent Communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 693–701.

Lu, C.; Kuba, J.; Letcher, A.; Metz, L.; Schroeder de Witt, C.; and Foerster, J. 2022. Discovered policy optimisation. *Advances in Neural Information Processing Systems*, 35: 16455–16468.

Micheli, V.; Alonso, E.; and Fleuret, F. 2023. Transformers are Sample-Efficient World Models. In *The Eleventh International Conference on Learning Representations*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540): 529–533.

Moeini, A.; Wang, J.; Beck, J.; Blaser, E.; Whiteson, S.; Chandra, R.; and Zhang, S. 2025. A survey of in-context reinforcement learning. *arXiv preprint arXiv:2502.07978*.

Pan, Y.; Li, T.; and Zhu, Q. 2025. Model-Agnostic Meta-Policy Optimization via Zeroth-Order Estimation: A Linear Quadratic Regulator Perspective. *arXiv preprint arXiv:2503.00385*. [Online] Available at `https://arxiv.org/pdf/2503.00385`.

Pignatelli, E.; Liesen, J.; Lange, R. T.; Lu, C.; Castro, P. S.; and Toni, L. 2024. NAVIX: Scaling MiniGrid Environments with JAX. *arXiv preprint arXiv:2407.19396*.

Raparthy, S. C.; Hambro, E.; Kirk, R.; Henaff, M.; and Raileanu, R. 2024. Generalization to New Sequential Decision Making Tasks with In-Context Learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML, 42138–42158.

Reed, S. 2022. A Generalist Agent. *Transactions on Machine Learning Research*.

Robine, J.; Höftmann, M.; Uelwer, T.; and Harmeling, S. 2023. Transformer-based World Models Are Happy With 100k Interactions. In *The Eleventh International Conference on Learning Representations*.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T.; and Silver, D. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609.

Schrittwieser, J.; Hubert, T.; Mandhane, A.; Barekatain, M.; Antonoglou, I.; and Silver, D. 2021. Online and Offline Reinforcement Learning by Planning with a Learned Model. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 27580–27591.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schwarzer, M.; Rajkumar, N.; Noukhovitch, M.; Anand, A.; Charlin, L.; Hjelm, R. D.; Bachman, P.; and Courville, A. C. 2021. Pretraining representations for data-efficient reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 12686–12699.

Stooke, A.; Lee, K.; Abbeel, P.; and Laskin, M. 2021. Decoupling Representation Learning from Reinforcement Learning. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 9870–9879. PMLR.

Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2016. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Toledo, E.; and Prorok, A. 2024. CoDreamer: Communication-Based Decentralised World Models. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Blaser, E. H.; Daneshmand, H.; and Zhang, S. 2024. Transformers Learn Temporal Difference Methods for In-Context Reinforcement Learning. In *ICML 2024 Workshop on In-Context Learning*.

Wang, J. X.; Kurth-Nelson, Z.; Tirumala, D.; Soyer, H.; Leibo, J. Z.; Munos, R.; Blundell, C.; Kumaran, D.; and Botvinick, M. 2016. Learning to reinforcement learn. *arXiv*.

Xu, M.; Shen, Y.; Zhang, S.; Lu, Y.; Zhao, D.; Tenenbaum, J.; and Gan, C. 2022. Prompting Decision Transformer for Few-Shot Policy Generalization. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 24631–24645. PMLR.

Yang, Y.-T.; Li, T.; and Zhu, Q. 2025. Transparent Tagging for Strategic Social Nudges on User-Generated Misinformation. *IEEE Transactions on Network Science and Engineering*, 1–14.

Zhu, C.; Dastani, M.; and Wang, S. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1): 4.

Appendix

A Code availability

The complete source code for CORAL and the experiments presented in this paper is available at https://github.com/fernando-ml/CORAL to ensure reproducibility. Our implementation is built in JAX (Bradbury et al. 2018) and leverages significant components, particularly for environment vectorization and training loops, from the PureJaxRL framework (Lu et al. 2022)[2] and Navix (Pignatelli et al. 2024)[3]. While core architectural choices (Appendix C) and most hyperparameters for baseline algorithms are consistent with PureJaxRL defaults for Navix.

B Limitations and Future Work

While our work demonstrates the significant potential of using pre-trained communicative priors for in-context adaptation, it also highlights several avenues for future extension.

A primary direction involves the **scalability and nature of the communication protocol itself**. The fixed-dimensional dense vector used in CORAL, while effective in the studied domains, may not be the most efficient representation for tasks with higher-dimensional inputs or those requiring more compositional reasoning. Future work could explore more structured protocols, such as discrete tokens from a learned vocabulary, to enhance expressivity. Furthermore, our framework operates in the "cheap talk" setting where communication is costless (Crawford and Sobel 1982; Lowe et al. 2019; Li and Zhu 2023; Yang, Li, and Zhu 2025). Introducing a cost for sending messages could be a valuable extension, potentially encouraging the IA to develop an even more efficient and sparse protocol where it communicates only when the expected utility outweighs the cost.

Another important avenue lies in **extending the validation of CORAL to more sophisticated domains.** Our experiments show successful transfer and adaptation within the *Navix* family of grid-worlds. An interesting next step is to evaluate the framework's ability to bridge a larger "transfer gap" by pre-training on one class of environments and deploying in fundamentally different ones, such as high-dimensional Atari games or continuous control tasks in MuJoCo. Success in such disparate domains would be a significant step towards truly generalist communicative agents.

Lastly, our framework can be extended to more **complex multi-agent topologies.** The current work deliberately focuses on an asymmetric agent pair to isolate and study the communicative prior. A natural extension is to scale this to scenarios where a single Information Agent must learn to coordinate multiple, potentially specialized, Control Agents. Future work could also explore settings with bi-directional communication, where all agents are homogeneous and can adopt both speaker and listener roles, moving towards to more applicable real-world coordination problems.

[2] https://github.com/luchris429/purejaxrl/tree/main

[3] https://github.com/epignatelli/navix

C Experiment Setup and Hyperparameters

Software We used the following software versions:

- Python 3.10.18 - Python Software License https://docs.python.org/3/license.html
- CUDA 12.4 - NVIDIA Software License Agreement https://docs.nvidia.com/cuda/eula/index.html
- Jax 0.5.3 - Apache License 2.0 https://github.com/jax-ml/jax
- Flashbax 0.1.2 - Apache License 2.0 https://github.com/instadeepai/flashbax
- Chex 0.1.89 - Apache License 2.0 https://github.com/google-deepmind/chex
- Optax 0.2.4 - Apache License 2.0 https://github.com/google-deepmind/optax
- Flax 0.10.4 - Apache License 2.0 https://github.com/google/flax
- Navix 0.7.0 - Apache License 2.0 https://github.com/epignatelli/navix
- WANDB 0.19.8 - MIT License https://github.com/wandb/wandb
- Gymnax 0.0.8 - Apache License 2.0 https://github.com/RobertTLange/gymnax
- PureJaxRL - Apache License 2.0 https://github.com/luchris429/purejaxrl

Hardware All experiments were conducted on NVIDIA Tesla V100-PCIE-32GB GPUs. A typical experimental run, consisting of training one algorithm over 30 random seeds for 10^7 total environment time steps on a Navix environment, completed in approximately 12 to 26 minutes.

C.1 Pre-training Environment Distribution

To foster the development of a generalizable transferable communication protocol, our pre-training stage utilizes a diverse distribution of tasks, $\mathcal{T}$. The environments were selected to expose the Information Agent (IA) to a wide range of mechanics, entities, and objectives. This multi-world regime is designed to drive the IA to learn a model of grid-world dynamics, rather than overfitting to the specifics of a single task. The pre-training distribution consists of the following environments:

- **Empty-Random-8x8**: The simplest task, it requires the agent to navigate to a goal in an empty room with randomized start and goal positions. It serves to teach the IA concepts of navigation and goal-directedness in the absence of other factors.
- **GoToDoor8x8**: This task requires the agent to navigate to a specific object (the door) rather than a generic goal tile. This teaches the IA the concept of object-centric goals so it can communicate information about specific entities in the environment.
- **FourRooms**: In this environment, the agent needs to navigate through multiple interconnected rooms. This tests the IA's ability to model and communicate information relevant to planning and navigation with multi-room layouts.

- **CrossingS9N3**: In this environment, the agent must cross a room with static obstacles, teaching the IA to learning about path planning and the semantics of impassable objects like walls.
- **LavaGapS6**: It introduces a "dangerous" terrain type (lava) that the agent must avoid, introducing IA to the concept of environmental constraints and affordances.
- **Dynamic-Obstacles-5x5**: This environment features mobile obstacles (balls) that move randomly. Since this environment is non-stationary, the IA must develop a compelling understanding of the world and communicate about changing states.
- **DoorKey-Random-6x6**: This task introduces the concept of conditional objectives. Here, the agent must first find a key, pick it up, and then use it to open a locked door to reach the goal. For this environment, the IA needs to model a longer sequence of sub-tasks and dependencies between them, and communicate accordingly.

C.2 Neural Network Architectures and Hyperparameters

CORAL's training and deployment setup and neural network designs are presented in Table 3, 4, and 5, respectively.

D Quantitative Analysis Details

D.1 Statistical Significance Testing

The performance metrics reported in Table 1 (Time-To-Threshold) and Table 2 (Zero-Shot Performance) are means computed over 30 independent seeds which are subject to statistical variability. To determine whether the observed differences between CORAL and the baseline methods (PPO, World Model) are statistically meaningful, we performed pairwise hypothesis testing. For each environment and each pair of methods (e.g., CORAL vs. PPO), we used an independent two-sample **Welch's t-test**. The test does not assume that the variance of outcomes is equal across different methods. We report a difference as statistically significant if the test yields a p-value below our chosen significance level of $\alpha = 0.05$.

All **confidence intervals** in the paper are 95% confidence intervals, calculated as $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{N}}$, where $\bar{x}$ is the sample mean, σ is the sample standard deviation, and $N = 30$ is the number of seeds.

D.2 Time-to-Threshold (TTT) Calculation

To provide a measure of sample efficiency that complements the learning curves, we use a Time-to-Threshold (TTT) analysis. TTT quantifies the number of environment steps required for an agent to reliably fulfill the task, or at least reach a high level of performance. Our calculation follows a two-pass process to ensure fair comparisons across environments with different reward scales and performance ceilings.

1. **Performance Thresholds**: Initially, for each evaluation environment, we determine the maximum asymptotic performance achieved across all runs of all methods. This sets an empirical "best-case" performance for the task. The performance threshold for each environment is then set to 90% of this maximum value.
2. **Calculating TTT and Success Rate**: In the second pass, for each individual seed of each method, we identify the first timestep at which the agent's episodic return meets or exceeds the calculated 90% performance threshold. This timestep is recorded as the TTT for that run. If a run fails to reach the threshold within the maximum allowed timesteps for that environment, its TTT is considered undefined. The final TTT reported in Table 1 is the mean over all successful runs. To capture the reliability of each method, we also present the Success Rate (SR), which is the percentage of the 30 independent runs that successfully reached the performance threshold.

Table 3: Train & Deployment Hyperparameters for CORAL. Shared hyperparameters between IA and CA come from Navix's PureJaxRL example for PPO.

Parameter	Value
Configuration	
Number of parallel environments (Pre-train)	128
Number of parallel environments (Deploy)	16
Total timesteps (Pre-train)	5×10^7
Total timesteps (Deploy)	1×10^7
Number of steps per rollout	128
Number of minibatches	8
Update epochs	4
Learning Parameters	
Optimizer	Adam
Learning rate (LR)	2.5×10^{-4}
Linear learning rate decay	True
Max gradient norm	0.5
Discount factor (γ)	0.99
GAE lambda (λ)	0.95
PPO clip epsilon	0.2
Loss Coefficients	
Entropy coefficient	0.02
Value function coefficient	0.5
Dyn. Awareness coefficient λ_{Dyn}	0.5
Causal coefficient λ_{Causal}	0.1
Temporal Coherence coefficient λ_{Coh}	0.05
Hybrid α for $\mathcal{U}_t$	0.5
Network Architecture	
Hidden dimension	128
Message dimension	32
Activation function	tanh
Transformer (Information Agent)	
Context length	4
Number of attention heads	4
Training Modes	
Pretrain mode	Both IA and CA trained
Deploy mode	IA frozen, CA re-trained

Table 4: Information Agent (IA) network architecture for CORAL. `Dense(in, out)` denotes a fully connected layer. The IA processes partial observations through a transformer architecture to generate communication messages and world model predictions.

CORAL Information Agent Architecture

Variables:
`obs_dim` = Partial observation vector dimension (environment-dependent)
`action_dim` = Number of discrete actions (7 for Navix environments)
`message_dim` = communication vector dimension
`hidden_dim` = latent representation dimension
`context_len` = transformer context buffer length
`num_heads` = self-attention heads
`activation` = tanh

Information Agent (IA) - Message Generation

```
▷ Transformer-based agent processing partial observations to generate messages.
 obs_tok = Dense(obs_dim, hidden_dim)
 pos_embed = LearnedParameter(context_len, hidden_dim)
 ln1, ln2 = LayerNorm(), LayerNorm()
 self_attn = SelfAttention(num_heads, hidden_dim)
 mlp_1 = Dense(hidden_dim, hidden_dim * 2)
 mlp_2 = Dense(hidden_dim * 2, hidden_dim)
 message_head = Dense(hidden_dim, message_dim)
```

IA Forward Pass (Message Generation):

```
obs_features = activation(obs_tok(partial_obs))
context_buf = concat([hidden_state[1:], obs_features[None]])
x = context_buf + pos_embed
y = self_attn(ln1(x)) + x   (residual connection)
z = mlp_2(activation(mlp_1(ln2(y)))) + y   (residual connection)
message = tanh(message_head(z[-1]))
return message, context_buf
```

Information Agent (IA) - World Model

```
▷ Additional prediction heads for dynamics modeling during training.
 next_obs_head = Dense(message_dim + action_dim, obs_dim)
 reward_head = Dense(message_dim + action_dim, 1)
 done_head = Dense(message_dim + action_dim, 1)
 next_msg_head = Dense(message_dim + action_dim, message_dim)
```

IA World Model Forward Pass:

```
message, context_buf = (as above)
world_input = concat([message, one_hot(action, action_dim)])
next_obs = clip(next_obs_head(world_input), -10.0, 10.0)
reward = clip(reward_head(world_input), -10.0, 10.0)
done = clip(sigmoid(done_head(world_input)), 1e-7, 1-1e-7)
next_message = tanh(next_msg_head(world_input))
```

Table 5: Control Agent (CA) network architecture for CORAL. `Dense(in, out)` denotes a fully connected layer. The CA processes both partial observations and messages from the IA to produce policy and value estimates.

CORAL Control Agent Architecture (Actor-Critic)

Variables:
`obs_dim` = Partial observation vector dimension (environment-dependent)
`action_dim` = Number of discrete actions (7 for Navix environments)
`message_dim` = Communication vector dimension
`hidden_dim` = Latent representation dimension
`activation` = tanh

Control Agent (CA)

```
▷ Processes partial observation and message to produce policy and value
estimates.
 obs_layer = Dense(obs_dim, hidden_dim)
 msg_layer = Dense(message_dim, hidden_dim)
 shared_1 = Dense(hidden_dim * 2, hidden_dim)
 shared_2 = Dense(hidden_dim, hidden_dim)
 actor_head = Dense(hidden_dim, action_dim)
 critic_head = Dense(hidden_dim, 1)
```

CA Forward Pass:

```
obs_features = activation(obs_layer(partial_obs))
msg_features = activation(msg_layer(message))
combined = concat([obs_features, msg_features])
x = activation(shared_1(combined))
x = activation(shared_2(x))
policy = Categorical(actor_head(x))
value = critic_head(x)
return policy, value
```

E Ablation Studies

E.1 Importance of Message Coherence

To validate the importance of using the Temporal Coherence Loss $\mathcal{L}_{\text{Coh}}$, we conducted an ablation study in the Information Agent's training. $\mathcal{L}_{\text{Coh}}$ is designed to promote temporal coherence in the emergent communication protocol, hypothesizing that a more predictable and consistent message stream brings a more stable learning signal for the Control Agent.

To test this, we pre-trained an IA without the $\mathcal{L}_{\text{Coh}}$ to observe its ability to guide a new CA. The results presented in Table 4 show that across all tested environments, the full CORAL agent consistently demonstrates superior sample efficiency and less variance compared to its variant without $\mathcal{L}_{\text{Coh}}$. The performance degradation is more noticeable in more complex tasks.

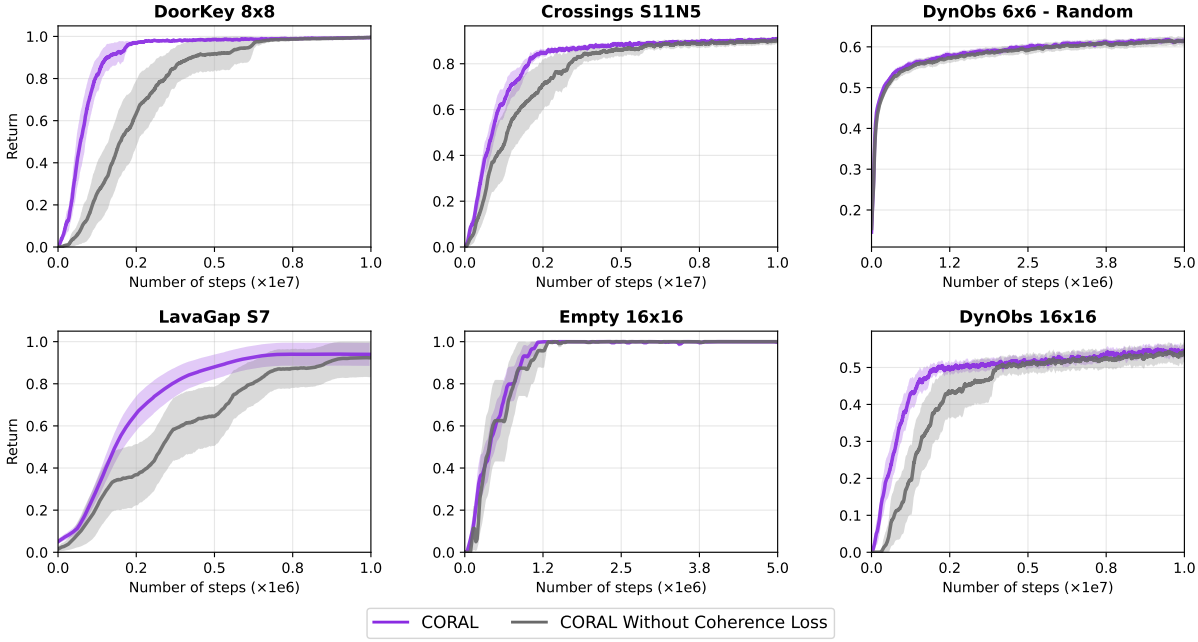

Figure 4: **Ablation study on the Message Coherence Loss** Learning curves show the mean episodic return ($\pm 95\%$ confidence interval) across 30 multiple seeds for a randomly initialized Control Agent paired with a pre-trained, frozen CORAL IA. CORAL demonstrates better sample efficiency and asymptotic performance compared to an ablated version where the Information Agent was pre-trained without the Temporal Coherence Objective. The consistent performance degradation in the ablated agent across unseen environments demonstrates that promoting temporal coherence in the communication protocol is critical for achieving better sample efficiency.

E.2 Information Agent Architecture

Our Information Agent's design is based on a Transformer architecture to model the history of observations. We conducted an ablation study comparing it against a more traditional recurrent alternative. We implemented and pre-trained a version of the CORAL Information Agent where the Transformer block was replaced by a Gated Recurrent Unit (GRU) cell (Chung et al. 2014). The GRU-based IA has comparable parameters and size, and is trained with the exact set of objectives as our main Transformer-based agent.

The in-context guided adaptation performance influenced by both architectures is presented in Table 5. The resulting learning curves show that CORAL Transformer outperforms or matches CORAL GRU in all environments.

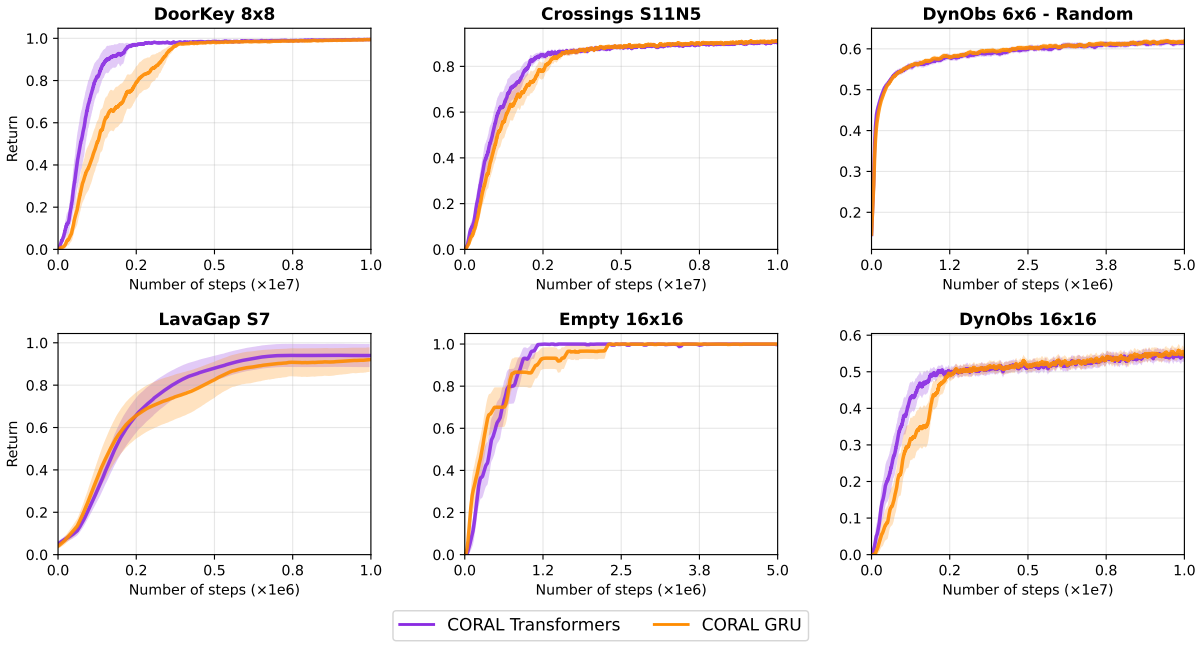

Figure 5: **Ablation study on the Information Agent's architecture.** Learning curves compare the in-context adaptation performance when the IA is implemented with a Transformer-based model versus a GRU-based alternative.

E.3 Message Dimensionality

We performed an ablation study to analyze the sensitivity of CORAL to the dimensionality of the message vector. We pre-trained and deploy three versions of the framework using different message sizes: 16, 32, and 64, while keeping all other hyperparameters unchanged.

Table 6 shows that CORAL is relatively robust to this choice, with all variants learning efficiently, specially compared to the baselines studied in our experimental results. **Message Size of 32** consistently achieves the best or near-best performance across the majority of tasks. The performance of the agent using a message size of 16 is often marginally slower, while the largest size of 64 also exhibits a minor lag in learning speeds in several environments. Empirically, a message size of 32 demonstrated the most consistent performance, justifying its use in our main experiments.

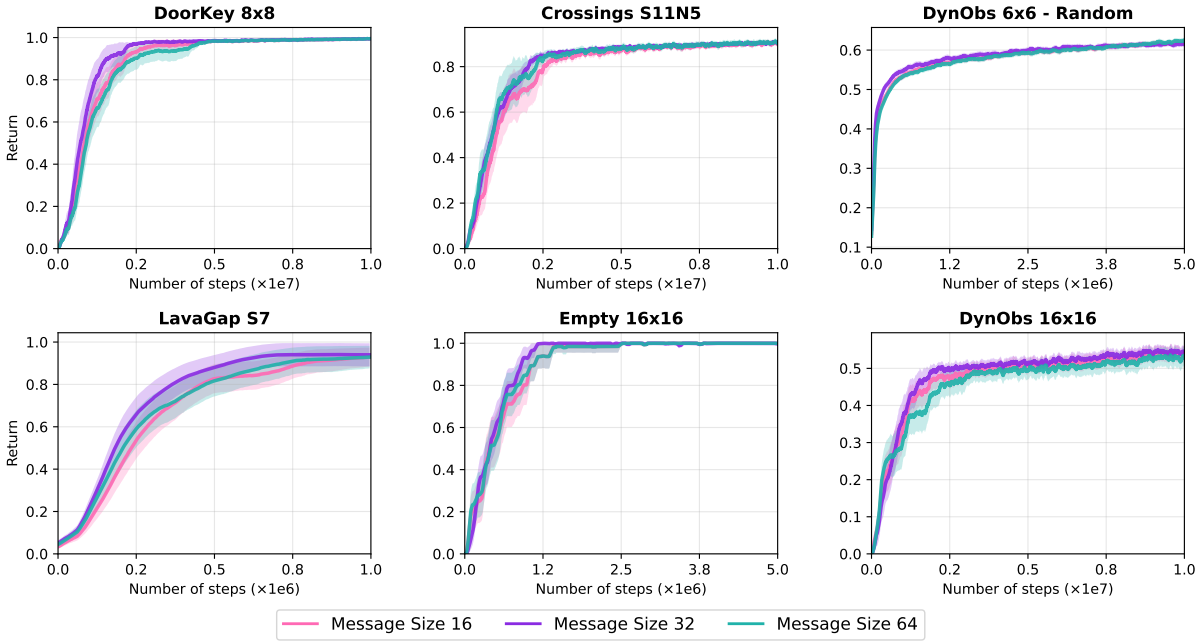

Figure 6: **Ablation study on message dimensionality.** In-context adaptation performance of CORAL with different message vector sizes (16, 32, 64). While CORAL is robust to this hyperparameter, a message of size 32 consistently provides a slight performance advantage.