CoACT: Coordination via Aligned Centralized Training in Multi-Agent Reinforcement Learning

Oussama Azizi[1,2], Frans A. Oliehoek[1], and Matthijs T.J. Spaan[1]

[1] Department of Intelligent Systems, Delft University of Technology
o.azizi@tudelft.nl
[2] Mercury Machine Learning ICAI Lab

Abstract. Cooperative Multi-Agent Reinforcement Learning (MARL) has achieved remarkable success in complex tasks, with Centralized Training Decentralized Execution (CTDE) being the dominant paradigm for training cooperative agents. However, most CTDE methods do not exploit the fact that agents are heterogeneous, each receiving observations generated by a different subset of state factors. In this work, we show that this heterogeneity can be exploited for a more efficient learning: by treating agents observation as partial views of the same underlying factors, and leveraging this structural dependency, we can can align the agents' representations during training to improve sample efficiency of existing state-of-the-art CTDE algorithms.

Keywords: Multi-Agent · Reinforcement Learning · Causal Inference · Representation learning.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) has gained popularity in the recent years in multiple domains ranging from cooperative-competitive real-time strategy games [28, 1], traffic control [29, 4], and robotics [11, 8]. A central challenge in scaling these methods to real-world settings is sample inefficiency, since learning effective policies require a prohibitive number of sample environment interactions, a problem that is exacerbated with high-dimensional observation spaces.

A natural approach to tackle this issue is to incorporate strong inductive biases into the learning process. One line of work [21, 24] does so by incorporating the reward structure of cooperative tasks by factorizing the global value function, reducing the complexity of the joint optimization problem. A complementary approach is to impose structure on how observations are processed and acted upon. Two prominent strategies have emerged: (1) One approach is to exploit symmetries in the state-action space by embedding directly equivariance constraints in the policy and value networks [20, 15]. (2) Another strategy, which has seen success in the single agent setting, is to use data augmentation and regularize the learned observation representations by encouraging invariance with respect to task-irrelevant information [13, 12, 10].

While representation learning has been extensively studied in the single agent setting, its principled application to MARL remains largely unexplored. A natural approach is to exploit an additional source of structure: multiple agents simultaneously observe the environment from a different view. This suggests treating the agents observations as different views of a shared underlying world state, and exploiting this redundancy in information across views to learn better representations. This idea has been exploited under the Centralized Training Decentralized Execution (CTDE) paradigm through a consensus mechanism, where agents explicitly align their representations during execution [30]. However, they treat the underlying state as a single monolithic block, implicitly assuming that all agents observe the same underlying factors. This ignores a fundamental characteristic of multi-agent systems, that is, agents are heterogeneous and can each receive observations that depend on different subsets of the environment's factors.

In this work, we propose to explicitly leverage this heterogeneity to improve sample efficiency, casting the problem as one of causal representation learning [31, 6], by considering each agent's observation as a partial view of the shared factors. First, we introduce *Structured-Observations Dec-POMDP* (SO Dec-POMDPs) an extension of the standard Dec-POMDP that explicitly models sparse, dynamic dependencies between state factors and agents' observations. Then, we leverage this structural dependency to define a principled representation objective, *Coordination via ALigned Centralized Training* (CoACT), that can be combined with any existing CTDE algorithm without modifying its architecture. Specifically, we encourage agents that share a common observed factor to produce consistent representations of that factor, while leaving agent-specific information intact. We provide the identifiability guarantees showing that, under a redundancy condition [6], this objective recovers the information of each shared latent factor. We further show empirically that adding this objective during training improves sample efficiency.

The rest of the paper is organized as follows. Section 2 introduces the necessary background on Dec-POMDPs and Contrastive Learning. Section 3 presents the proposed Structured-Observations Dec-POMDP framework and its related assumptions. Section 4.1 introduces the CoACT loss and Section 4.2 provides its identifiability guarantees. Section 5 empirically demonstrates the benefits of CoACT in terms of sample efficiency in various environments. Section 6 reviews related work. Finally, Section 7 concludes with a discussion, limitations, and future work.

2 Background

2.1 Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs)

A Dec-POMDP [18] is a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, T, \mathcal{R}, \Omega, \mathcal{O}, b_0, \gamma, H \rangle$, where $\mathcal{I} = \{1, \dots, N\}$ is a set of finite agents, $\mathcal{S}$ is the global state, $\mathcal{A} = \prod_{i \in \mathcal{I}} \mathcal{A}_i$ is the joint action space, with $\mathcal{A}_i$ the action space of agent i, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is

the transition kernel $T(s'|s,a)$, $\Omega = \prod_{i\in\mathcal{I}} \Omega_i$ is the joint observation space, $\mathcal{O} : \mathcal{S} \times \Omega \to [0,1]$ the observation kernel, $\mathcal{O}(o|s)$, $b_0 \in \Delta(\mathcal{S})$ the initial state distribution, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the shared reward function, $\gamma \in [0,1)$ the discount factor, and H the horizon. At each time step t, given the global state s_t, the agents $\mathcal{I}$ receive joint observations $o_t = (o_{1,t}, \dots o_{N,t}) \sim \mathcal{O}(.|s_t)$. Each agent i takes an action $a_{i,t} \in \mathcal{A}_i$, based on its history $h_{i,t} = [o_{i,0}, a_{i,0}, \dots, a_{i,t-1}, o_{i,t}]$, yielding a joint action a_t. The environment returns a reward $r_t = R(s_t, a_t)$ and transitions to the next state $s_{t+1} \in \mathcal{S}$ according to the transition probability $T(s_{t+1}|s_t, a_t)$. The goal is to find a joint policy $\pi = \{\pi_i\}_{i\in\mathcal{I}}$ that maximizes the expected discounted return $\mathbb{E}[\sum_{t=0}^{H} \gamma^t r_t]$. When the observations are high-dimensional, each policy π_i is parameterized by neural-networks, by first encoding each observation into a representation $z_{i,t} = g_{\psi_i}(o_{i,t})$. Then a history-aware representation $f_{\phi_i}(z_{i,0}, \dots z_{i,t})$ is constructed using for instance Long Short Term Memory (LSTM) Networks [9], which are used to sample the actions.

Centralized Training Decentralized Execution (CTDE): CTDE is a multi-agent training paradigm which addresses the partial observability and non-stationarity that can arise in MARL. During training, CTDE allows agents to have access to the full global information, often through a centralized critic, while restricting the execution to the local observations.

2.2 Contrastive Learning

Contrastive Learning is a self-supervised learning technique that aims at training an encoder $g : \mathcal{O} \to \mathcal{Z}$ mapping observations o to representations z that are maximally informative about their generative factors without having access to the ground truth [7, 2, 19]. Specifically, most prominent Contrastive Learning losses are defined in terms of an *anchor* o, a *positive* sample o^+, and a set of L *negative* samples $\{o_l^-\}_{l=1}^L$ which are assumed to be generated by different latent factors. These losses are expressed as (1) an alignment term that maximizes the similarity between the *anchor* and *positive* sample (2) a uniformity or contrastive term that aims at minimizing the similarity between the *anchor* and the *negative* samples. For instance, the SimCLR loss [2] is expressed as:

$$\mathcal{L}_{\text{SimCLR}}(g) = \mathbb{E}_{o,o^+,\{o_l^-\}}\Big[\log\Big[\frac{\exp(\text{sim}(g(o), g(o^+))/\tau)}{\sum_{l=1}^{L}\exp(\text{sim}(g(o), g(o_l^-))/\tau)}\Big]\Big], \tag{1}$$

where $\text{sim}(u, v)$ is the cosine similarity between u and v.

3 Structured-Observations Dec-POMDPs

While standard Dec-POMDPs assume a monolithic state space and observation functions, many real-world multi-agent applications exhibit spatio-temporal structure, where the observations depend dynamically on a sparse set of causal latents. For instance, in traffic control [29, 4], each agent observes only nearby

intersections, and in real-time strategy games [28, 1] units are constrained by a limited field of view. Motivated by such real-world applications, we define the Structured-Observations Dec-POMDP framework which is able to capture these causal dependencies.

A Structured-Observations Dec-POMDP is defined as a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, T, R, \Omega, \mathcal{O}, \mathcal{G}, b_0, \gamma, H \rangle$. This formalism is identical to a standard Dec-POMDP, with the exception of the following components which explicitly define the structural dependence between the state and the observations:

1. The state space $\mathcal{S} = \prod_k^K \mathcal{S}_k$, which is factorized into K components,
2. The state-observation dependency mapping $\mathcal{G} : \mathcal{S} \to \mathcal{P}(1, \ldots, K)^{|\mathcal{I}|}$, which assigns for any agent $i \in \mathcal{I}$ the subset of state factors K_i influencing its observation given the state s such that $\mathcal{G}(s) = \{K_i\}_{i \in \mathcal{I}}$,
3. The observation function which is factorized following the state-observation dependency as: $\mathcal{O}(o_t|s_t) = \prod_{i \in \mathcal{I}} \mathcal{O}_i(o_{i,t}|\{s_{k,t} \mid k \in K_{i,t}\})$.

Consider for instance the Gridworld environment depicted on the left in Figure 1, where two agents navigate in a grid where they need to collect an object and move towards a target destination. The limited field of view of each agent is only capable of capturing information about objects in its immediate vicinity; as the agent navigates the environment, this set of causally relevant objects changes over time. In this sense, our framework generalizes the standard Dec-POMDP formulation, allowing for such dynamic shifts, without excluding static structures where the state-observation dependency would be fixed for all states. The central challenge is to map high-dimensional observations, that mix information from multiple factors, to representations that isolate the information of each factor, in a way that is useful for the downstream Reinforcement Learning task.

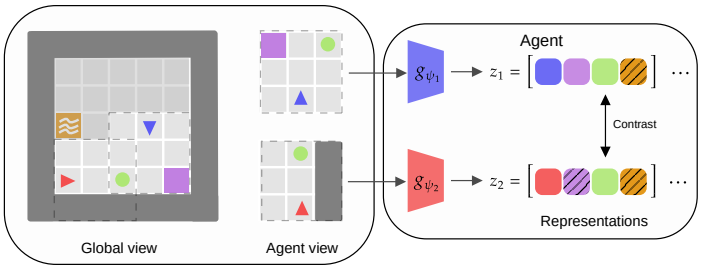


Fig. 1: Gridworld environment with two agents. On the left: the environment global view with the agents local views. The blue agent observes the target destination and the ball, the red agent observes only the ball, while both do not observe the lava field. On the right: The representation network of each agent encodes its observation into a factored representation corresponding each to one object, and an additional private factor (blue or red block), which encodes all the unique information about its observation. The hatched representation factors are masked, reflecting unobserved objects in the local views.

4 CoACT: Multi-Agent Coordination via Representation Alignment

To achieve the goal of learning causally-grounded representations which capture this underlying structure of the environment, we formalize in this section our representation learning objective. This objective is designated to leverage the state-observation dependency $\mathcal{G}$. Unlike standard CTDE methods that require access to the full global state, our approach relaxes this assumption by relying solely on the structure of the observation function.

4.1 The representation learning objective

In our framework, we assume having access to the state-observation mapping $\mathcal{G}(s) = [K_1, \dots K_{|\mathcal{I}|}]$ at every timestep. This mapping defines the subset of latents $K_i \subset K$ influencing the observation o_i of every agent $i \in \mathcal{I}$. From this evaluation, we dynamically derive the set of agents $\mathcal{I}_k$ associated with the latent k:

$$\mathcal{I}_k = \{i \in \mathcal{I} \mid k \in K_i\}. \tag{2}$$

Each agent i maps, through its representation network g_{ψ_i}, the local observation o_i to a factored representation $z_i = [z_{i,1}, \dots, z_{i,K}]$, capturing each their associated latent factor. For instance, in the aforementioned gridworld example in Figure 1, each agent representation consists of factors relating to one specific object in the environment in addition to one private factor.

We denote by $g_{\psi_{i,k}}$ the functional mapping from observations to the k-th representation factor, where $\psi_{i,k} \subset \psi_i$ is the subset of the parameters in the

computational graph of the representation network influencing the representation factor z_k. Formally $z_k = g_{\psi_{i,k}}(o_i)$. Note that $\psi_{i,k}$ and $\psi_{i,k'}$ may overlap due to shared parameters.

Given the joint observations $(o_1, \ldots, o_{|\mathcal{I}_k|})$, the *factor alignment objective* is computed over a batch of B by leveraging a pairwise contrastive loss, for instance the SimCLR loss:

$$\mathcal{L}_{\text{rep},k} = \frac{1}{2} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} \Big[\mathcal{L}_{\text{simCLR}}(g_{\psi_{i,k}}, g_{\psi_{j,k}}; \{o_{i,t}, o_{j,t}\}_{t=1}^B) + \mathcal{L}_{\text{simCLR}}(g_{\psi_{j,k}}, g_{\psi_{i,k}}; \{o_{i,t}, o_{j,t}\}_{t=1}^B) \Big], \tag{3}$$

where $\mathcal{L}_{\text{simCLR}}(g_{\psi_{i,k}}, g_{\psi_{j,k}}; \{o_{i,t}, o_{j,t}\}_{t=1}^B)$ is the empirical estimate of the SimCLR loss over the data $\{o_{i,t}, o_{j,t}\}_{t=1}^B$:

$$\mathcal{L}_{\text{simCLR}}(g_{\psi_{i,k}}, g_{\psi_{j,k}}; \{o_{i,t}, o_{j,t}\}_{t=1}^B) = \frac{1}{B} \sum_{t=1}^{B} \log \frac{\exp(\text{sim}(g_{\psi_i,k}(o_{i,t}), g_{\psi_j,k}(o_{j,t}))/\tau)}{\sum_{l=1}^{B} \exp(\text{sim}(g_{\psi_i,k}(o_{i,t}), g_{\psi_j,k}(o_{j,l}))/\tau)}. \tag{4}$$

The symmetry in the *factor alignment objective* ensures that the representation networks contribute equally as a source of negative samples, guaranteeing balanced gradient updates.

The CoACT representation learning objective is defined as the weighted average over all the *factors alignment objectives*:

$$\mathcal{L}_{\text{CoACT}} = \sum_{\substack{k \in \{1,\ldots,K\} \\ |\mathcal{I}_k| \geq 2}} \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \mathcal{L}_{\text{rep},k}, \tag{5}$$

which can be optimized in addition to the RL objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{RL}} + \alpha \mathcal{L}_{\text{CoACT}}. \tag{6}$$

This objective can be integrated into the policy representations or the value networks. Additionally, the global critic representations can also be aligned with the representations of the individual policies.

4.2 Identifiability guarantees

In this section, we show that using the aforementioned objective guides each representation factor towards extracting information of its respective latent. In the following, we drop the time dependency. We assume the following:

Assumption 1 (Conditional Independence of the observations) *The observations of agents $\mathcal{I}_k$ are conditionally independent given the shared latent:*

$$O_i \perp\!\!\!\perp O_j \mid S_k, \tag{7}$$

Additionally, we make the following information-theoretic assumption, adapted from [6] to the state-factor k:

Assumption 2 (Redundancy of information) *For any $(i, j) \in \mathcal{I}_k \times \mathcal{I}_k$, we have:*

$$I(O_i; S_k | O_j) = 0, \tag{8}$$

Namely, the observations of two distinct agents influenced by the shared factor k contain exactly the same information about it, or in other words, they are maximally redundant. Note that this is an information-theoretic assumption: it does not constrain the observation spaces themselves and only concerns redundancy on shared factors. The heterogeneity in the state-observation dependencies is still captured through $\mathcal{G}$. Given these assumptions, we can state our main identifiability guarantee:

Theorem 3 (Identifiability of the latent). *Given Assumptions 1 and 2 and the use of SimCLR loss as a pairwise alignment objective for the representation factors with L negative samples and a temperature $\tau = 1$, we have the following inequality:*

$$\frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k|-1)}\mathcal{L}_{rep,k} \geq \log(L) - \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k|-1)} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} [\underbrace{I(Z_{i,k}; S_k)}_{(a)} - \underbrace{I(Z_{i,k}; S_k | Z_{j,k})}_{(b)}]. \tag{9}$$

The proof is available in Appendix B. Theorem 3 states that the *factor alignment objective* is an upper bound on an information bottleneck over the representation which (a) encourages the recovery of the of latent state information, while (b) maintaining information overlap across agents' representation. Assumption 2 guarantees that asymptotically no state-information is discarded while the term (b) is minimized.

Remark 1. While Assumption 2 may be difficult to satisfy in practice, its violation merely implies that the representation objective will lead to a partial recovery of the state factor information. Even in such cases, the CoACT regularization remains beneficial during early stages of training, by guiding the model towards discarding (partially) state-irrelevant information. The hyperparameter α defined in Equation 6, can be subjected to a decay schedule to prioritize the RL objective at later stages of training.

5 Experiments

In this section, we showcase our approach in two environments characterized by high-dimensional image observations both with a static (Section 5.1) and dynamic (Section 5.2) state-observation dependency.

5.1 Double Digit prediction

The environment consists of $N = 3$ agents acting simultaneously on a shared pool of $N + 1$ digits sampled from the MNIST dataset [5] at the start of the episode.

The environment's horizon is $H = 1$.

Observations: At each timestep, agent i receives a 56×56 grayscale image observation constructed by placing two digits side by side, where the digit i is on the left and digit $i + 1$ is on the right, making each agent share effectively one latent with each one its neighbors as depicted in Figure 2a. The observations are generated each timestep by augmenting the same MNIST digit image using random affine transformations, distortions and random pixel masking.

Actions and rewards: Each agent selects a two-digit integer between 0 and 99, encoding the prediction for the observed digits. For each correctly predicted digit, the agent receives reward 1/2 and 0 otherwise, yielding a maximum reward of 1.

By construction, the environment provides a ground-truth structure where the digits act as the latent factors. The observations are arranged in a chain, such that adjacent agents share one latent factor.

5.2 Gridworld

We introduce a variant of the single agent minigrid environment [3] to account for multiple agents. In this environment, $N = 3$ agents need to navigate a 8×8 grid for $H = 100$ steps to collect $K = 5$ colored balls placed in random locations as shown in Figure 2b. The agents spawn at random locations and directions at the start of every episode.

Observations: Each agent observes 42×42 image corresponding to 7×7 tile window centered in front of the agent, which reflects faithfully what the agent can see from its position and orientation.

Actions and rewards: Each agent selects from 7 discrete actions: `turn left`, `turn right`, `move forward`, `pick up`, `drop`, `toggle` and `done`. The rewards are sparse, the agents receive a reward 10 for each picked ball and 0 otherwise.

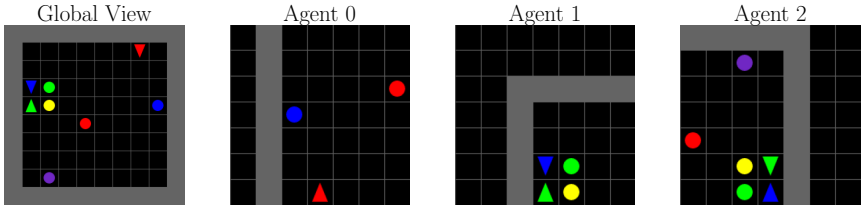
Causal mask: At every timestep, the environment additionally emits a causal mask for each agent, describing which objects (lava, ball, door, key, etc.) are inside the field of view of the agent.

For the Gridworld experiment, we use Multi-Agent PPO [32] with General Advantage Estimation and Popart normalisation. For the Digit Prediction experiment, we use IQL [25].

In both experiments we use the SimCLR loss with a temperature of 0.3 and $\alpha = 0.5$. For more details, we refer the reader to Appendix A.

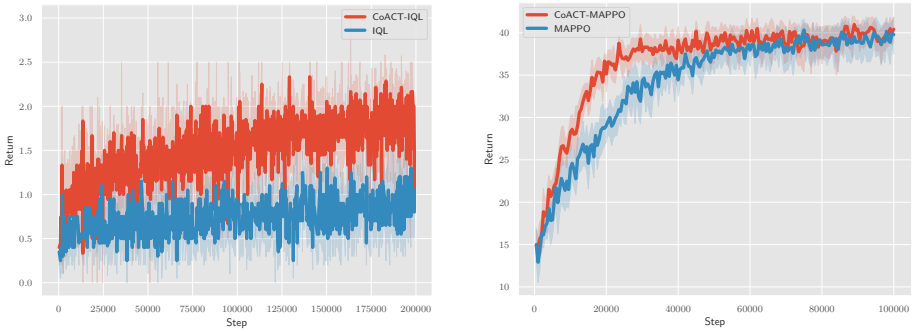
We run experiments on 10 seeds and report in Figures 3a and 3b the average and standard error of the returns over $200k$ environment steps for the Digit Prediction environment and $100k$ steps for the Gridworld environment.

(a) Digit prediction



(b) Gridworld

Fig. 2: Observation samples from the environments.



(a) Digit Prediction (b) Gridworld

Fig. 3: Training curves of the experiments averaged over 10 seeds.

Discussion: CoACT-MAPPO consistently outperforms MAPPO in the Gridworld environment, achieving higher returns earlier in training while converging to similar final performance, demonstrating a clear sample efficiency gain. In the Digit Prediction environment, both CoACT-IQL and IQL exhibit high variance across seeds, attributable to the stochastic nature of the observation augmentations. Nevertheless, CoACT-IQL maintains a consistent advantage in expected return throughout training.

6 Related work

6.1 Centralized Training Decentralized Execution (CTDE)

CTDE is a common training paradigm in cooperative Multi-Agent Reinforcement Learning methods. These methods enable the agents to leverage global information while maintaining decentralized execution. Methods such as QMIX [21], MAPPO [32], VDN [24], and MADDPG [14] demonstrated strong empirical performance across a variety of mixed cooperative-competitive tasks such as MPE [17] and SMAC [28]. However, these methods typically assume access to the global state during training, which is often unavailable or expensive to obtain in practice. In our work, we only assume access to the state-observation dependency $\mathcal{G}(s)$, which is strictly a weaker requirement than the full global state.

6.2 Representation Learning in RL

Representation Learning has emerged as a prominent subfield in Reinforcement Leaning with Deep RL methods [27, 22, 16]. Learning compact representations that compress high-dimensional observations in a meaningful way is crucial for the agent's ability to generalize and learn efficiently. A prominent line of work achieves this purpose either by means of Data-Augmentation without any auxiliary objective [13], or through a contrastive learning loss combined with Data-Augmentation [12]. These methods demonstrate significant improvement in the single agent setting. Similar ideas have been leveraged in the Multi-Agent setting, by leveraging a masked attentive contrastive [23], or using a shared discrete consensus representation across agents by treating their observations as different views of the same global state [30]. However, they treat the state as a monolithic block without considering possible heterogeneity in the agents' observations.

6.3 Multi-view representation learning

Multi-view representation learning aims at learning representations that capture information shared across views of the same underlying generating factor. [26] propose Contrastive Multiview Coding to learn view-invariant representations, by maximizing the Mutual-Information between across views, and show that leveraging more views improve the representations. [6] propose a theoretical framework through the Information Bottleneck principle. They introduce the notion of redundancy, i.e., that two views introduce the same task-relevant information. They show that maximizing the shared information, while discarding view-specific information leads to robust representations. [31] study identifiability guarantees where each view depends on a different subset of latent factors. In our work, we build on the redundancy notion introduced by [6] to derive identifiability guarantees of the latent factors.

7 Conclusion

In this work, we introduced CoACT, a modular representation learning objective for cooperative MARL that exploits the heterogeneous structure of agents' observations. We formalized the Structured-Observations Dec-POMDP to explicitly capture sparse dependencies between state factors and agent observations, and proposed a contrastive alignment objective that encourages agents observing the same latent factor to produce consistent representations. We provided theoretical guarantees showing that this objective recovers shared latent factor information under a maximal redundancy condition, and demonstrated empirically that CoACT improves sample efficiency when combined with both IQL and MAPPO, without modifying their architectures or requiring access to the full global state.

Limitations and future work: CoACT currently relies on the state-observation dependency at training time. Learning this structure from data is an important direction for future work. Several ablations also remain to be conducted, including a comparison against monolithic representation alignment, a study of alternative contrastive losses, and validation on larger benchmarks such as SMAC.

Acknowledgments. This research has been conducted in the Mercury Machine Learning Lab, and we express our gratitude to Booking.com for generously funding this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Implementation details

A.1 IQL

Encoder: A three-layer CNN with channels $(32, 64, 128)$, kernel size 6, stride 1, and padding 3, followed by ReLU activations. The CNN output is projected to a factored representation of dimension $3 \times 32 = 96$ (3 agents, 32 units per factor). Dropout $p = 0.1$ is applied after the projection.

Training: Each agent maintains an independent Q-network and replay buffer (capacity 50,000) with ϵ-greedy exploration ($\epsilon_0 = 0.999$, $\epsilon_{\min} = 0.01$, decay $= 0.995$ per episode). Gradients are computed on mini-batches of size 64 sampled from the replay buffer. The actor and critic learning rate is 10^{-3}, the discount factor is set at $\gamma = 1.$, and soft target-network updates use $\tau = 0.01$. Training runs for 200,000 episodes.

Q-network: A two-layer MLP with hidden size 64 and ReLU activations.

CoACT: The contrastive representation alignment loss uses temperature $\tau = 0.3$, coefficient $\alpha = 0.5$. The causal mask follows a chained structure over the 64-dimensional representation space: agent 0 attends to dims 0–31, agent 1 to dims 16–47, and agent 2 to dims 32–63. Adjacent agents share a 16-dimensional overlap (agents 0–1 share dims 16–31; agents 1–2 share dims 32–47), while non-adjacent agents 0 and 2 have disjoint factors.

A.2 MAPPO

Encoder: A three-layer CNN with channels $(16, 32, 32)$, kernels $(3, 3, 3)$, strides $(2, 2, 2)$, padding $(1, 1, 1)$, ReLU activations, and global average pooling. Inactive factor slots are zeroed (factor masking). Representations are factored into $F = 8$ factors of dimension 11, giving $\mathbf{z} \in \mathbb{R}^{88}$

Policy / Value networks: Actor and critic are two-layer MLPs with hidden size 64 and ReLU activations. The centralised critic receives concatenated observations of all agents.

CoACT: We pick the values of temperature $\tau = 0.3$, and $\alpha = 0.5$. Furthermore we use align the centralized critic as well with an alignment coefficient 0.2. Furthermore we use the MAPPO training parameters described in Table 1.

Parameter	Value
Actor LR	5×10^{-4}
Critic LR	10^{-3}
Discount γ	0.99
GAE λ	0.95
PPO clip ϵ	0.1
PPO epochs	5
Num. mini batch	4
Value loss coef.	1.0
Entropy coef.	10^{-3}
Max grad norm	5.0
Episodes / update	5
Total episodes	$20,000$

Table 1: Training parameters for MAPPO

B Missing proofs

We now provide the full proof of Theorem 3. We first introduce the necessary notation. Let O_i denote the random variable corresponding to the observation of agent i, and let S_k denote the random variable corresponding to the k-th latent factor. Each agent i maps its observation O_i to a factored representation $Z_i = g_{\psi_i}(O_i)$, where $Z_{i,k} = g_{\psi_{i,k}}(O_i)$ denotes the k-th factor of the representation. We use $I(\cdot;\cdot)$ to denote mutual information and $I(\cdot;\cdot|\cdot)$ to denote conditional mutual information.

The proof proceeds in three steps:

1. **(Section B.1.1)** We show that under Assumptions 4 and 5, maximizing the pairwise mutual information $I(O_i; O_j)$ is equivalent to maximizing the joint mutual information $I(O_i; \{O_j\}_{j \in \mathcal{I}_k \setminus \{i\}})$, reducing the problem to pairwise alignment.
2. **(Section B.1.2)** We show that the conditional independence of observations induces conditional independence of representations via the data processing inequality, i.e. $Z_{i,k} \perp\!\!\!\perp Z_{j,k} \mid S_k$.
3. **(Section B.1.3)** We show that minimizing the *factor alignment objective* loss provides a lower bound on $I(Z_{i,k}; S_k) - I(Z_{i,k}; S_k | Z_{j,k})$, guaranteeing recovery of the shared latent factor information, while maintaining information overlap between all agents representations during training.

B.1 Proof of Theorem 3

In this section we prove individual representation factor loss recovers the information related to the shared latent. In the following, let k in $\{1, \dots, K\}$ and assume $|\mathcal{I}_k| \geq 2$. We start by reminding the following assumptions, which are necessary to recover only the information about the targeted latent.

Assumption 4 (Joint conditional independence) *We assume we have the following conditional independence for all i in $\mathcal{I}_k$:*

$$O_i \perp\!\!\!\perp O_{\mathcal{I}_k \setminus \{i\}} | S_k \tag{10}$$

Remark 2. (Pairwise conditional independence) Assumption 4 implies the following independence for all $(i, j) \in \mathcal{I}_k \times \mathcal{I}_k$:

$$O_i \perp\!\!\!\perp O_j | S_k \tag{11}$$

Our objective is to recover the information of the latent S_k. This information is simply the overlap of information between the observations of all agents in $\mathcal{I}_k$. To extract this common information, we can maximize the joint mutual information between the kth representation factor of an agent $i \in \mathcal{I}_k$, $Z_{i,k}$ and the other agents in $\mathcal{I}_k$ $\{Z_{j,k} = g_{\psi_j}(O_j)\}_{j \in \mathcal{I}_k \setminus \{i\}}$. This quantity can be decomposed via the chain rule of information as:

$$I(Z_{i,k}; \{Z_{j,k}\}_{j \in \mathcal{I}_k \setminus \{i\}}) = \sum_{j \in \mathcal{I}_k \setminus \{i\}} I(Z_{i,k}; Z_{j,k} | Z_{1,k} ... Z_{j-1,k}) \tag{12}$$

While we could in principle aim to maximize the LHS, it poses architectural and, or computational complexities, since it requires a choice of an aggregation function accross $|\mathcal{I}_k| - 1$ views. Similarily, maximizing the RHS poses similar architectural complexities that do not scale when $\mathcal{I}_k$ is large. It essentially requires to incrementally recover overlaps of information. We make further assumptions to get rid of the high-order interactions. That is, we aim to have:

$$I(Z_{i,k}; \{Z_{j,k}\}_{j \in \mathcal{I}_k \setminus \{i\}}) = I(Z_{i,k}; Z_{j,k}) \tag{13}$$

This guarantees recovering the shared latent using existing contrastive learning methods.

B.1.1 Reducing joint to pairwise alignment We need to prove that minimizing the pairwise losses is equivalent to maximizing the joint mutual information between representations. This equivalence is necessary to leverage contrastive learning losses. We make the following information-theoretic assumption:

Assumption 5 (Redundancy of information [6]) *For any $(i, j) \in \mathcal{I}_k \times \mathcal{I}_k$, we have the following:*

$$I(O_i; S_k | O_j) = 0 \tag{14}$$

This assumption states that the information about the factor S_k is maximally redundant across the agents' observations. To see this, we can write the joint mutual information between O_i and (S_k, O_j) as follows:

$$I(O_i; S_k, O_j) = I(O_i; S_k) + I(O_i; O_j | S_k) \tag{15}$$

$$= I(O_i; O_j) + I(O_i; S_k | O_j) \tag{16}$$

Therefore, the MI between the observations is decomposed into two main terms:

$$I(O_i;O_j) = \underbrace{I(O_i;S_k) - I(O_i;S_k|O_j)}_{\text{(1) relevant}} + \underbrace{I(O_i;O_j|S_k)}_{\text{(2) irrelevant}} \tag{17}$$

The first term (1) is all the shared information relevant to the state S_k, which is simply the difference between the information about S_k contained in O_i and its unique information. While the second term (2) captures everything that is shared between O_i and O_j and is not related to S_k, which is zero by Assumption 4. Hence, we can write:

$$I(O_i;O_j) = I(O_i;S_k) - I(O_i;S_k|O_j) \geq 0 \tag{18}$$

Therefore, we have the following inequality, due to the positivity of the mutual information:

$$\underset{\text{maximum redundancy}}{0} \leq I(O_i;S_k|O_j) \leq \underset{\text{maximum uniqueness}}{I(O_i;S_k)} \tag{19}$$

When the uniqueness is maximal, (or to put it differently, the redundancy is minimal), that is $I(O_i;S_k|O_j) = I(O_i;S_k)$, we have $I(O_i;O_j) = 0$. This means that maximizing the mutual information between the observations can only destroy information about the state history S_k. Aligning representations in this case is harmful for the downstream task. When the uniqueness is minimal (or to put it differently, the redundancy is maximal), that is $I(O_i;S_k|O_j) = 0$, then:

$$I(O_i;O_j) = I(O_i;S_k), \tag{20}$$

which guarantees to recover all the information about S_k contained in the observations.

We can now prove the equivalence between the joint and pairwise MIs:

Theorem 6 (Equivalence between joint and pairwise MIs). *Given assumptions 4 and 5, we have for all $(i,j) \in \mathcal{I}_k \times \mathcal{I}_k$:*

$$I(O_i;\{O_{j'}\}_{j' \in \mathcal{I}_k \setminus \{i\}}) = I(O_i;O_j) \tag{21}$$

Proof. The conditional independence given by Assumption 4 implies that:

$$\begin{aligned} 0 &= I(O_i;O_{\mathcal{I}_k \setminus \{i\}}|S_k) && (22)\\ &= I(O_i;S_k,O_{\mathcal{I}_k \setminus \{i\}}) - I(O_i;S_k) && (23)\\ &= I(O_i;O_{\mathcal{I}_k \setminus \{i\}}) + I(O_i;S_k|O_{\mathcal{I}_k \setminus \{i\}}) - I(O_i;S_k) && (24) \end{aligned}$$

We show that $I(O_i;S_k|O_{\mathcal{I}_k \setminus \{i\}}) = 0$:

$$I(O_i;S_k,O_{\mathcal{I}_k \setminus \{i,j\}}|O_j) = \underbrace{I(O_i;S_k|O_j)}_{=0 \text{ by Assumption 5}} + I(O_i;O_{\mathcal{I}_k \setminus \{i,j\}}|O_j,S_k) \tag{25}$$

$$\tag{26}$$

And by Assumption 4 and its implied pairwise conditional independence in Remark 2 we have $I(O_i;O_{\mathcal{I}_k \setminus \{i,j\}}|O_j,S_k) = I(O_i;O_{\mathcal{I}_k \setminus \{i\}}|S_k) - I(O_i;O_j|S_k) = 0$.

Therefore we have from Equation (24) and (20) that $I(O_{i,t};O_{\mathcal{I}_k \setminus \{i\}}) = I(O_i;S_k) = I(O_i;O_j)$ for all $j \in \mathcal{I}_k \setminus \{i\}$.

B.1.2 Conditional independence transfers to representations Having established that pairwise alignment suffices, we now lift the conditional independence from observations to representations. Since representations are deterministic functions of observations, the data processing inequality implies that the conditional independence structure is preserved:

Lemma 1 (Conditional independence of representations). *Under Assumption 4, for all $(i, j) \in \mathcal{I}_k \times \mathcal{I}_k$, we have:*

$$Z_{i,k} \perp\!\!\!\perp Z_{j,k} \mid S_k$$

Proof. Since representations are deterministic functions of observations, applying the data processing inequality twice yields:

$$0 = I(O_i; O_j \mid S_k) \geq I(Z_{i,k}; O_j \mid S_k) \geq I(Z_{i,k}; Z_{j,k} \mid S_k)$$

By positivity of mutual information, $I(Z_{i,k}; Z_{j,k} \mid S_k) = 0$, hence $Z_{i,k} \perp\!\!\!\perp Z_{j,k} \mid S_k$.

Remark 3. Analogously to Equation 15, and using the conditional independence established in Lemma 1, the mutual information between pairs of representations decomposes as:

$$I(Z_{i,k}; Z_{j,k}) = I(Z_{i,k}; S_k) - I(Z_{i,k}; S_k \mid Z_{j,k}) \tag{27}$$

B.1.3 Latent recovery via contrastive learning Having established that $I(Z_{i,k}; Z_{j,k})$ decomposes into identifiability-relevant terms via Equation 27, it remains to connect the SimCLR loss to this mutual information. We leverage the following result from [19], which shows that the SimCLR loss, with a temperature $\tau = 1$ provides a lower bound on the mutual information between pairs of representations, with the bound becoming tighter as the number of negative samples L increases:

Lemma 2 (Mutual Information lower bound [19]).

$$I(Z_{i,k}; Z_{j,k}) \geq \log(L) - \mathcal{L}_{simCLR}(g_{\psi_{i,k}}, g_{\psi_{j,k}}) \tag{28}$$

Applying Lemma 2 to each pairwise term in the factor alignment objective and substituting the decomposition from Equation 27 yields the following result. The $\frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k|-1)}$ weighting normalizes over all ordered pairs in $\mathcal{I}_k$, ensuring each agent contributes equally regardless of the number of agents observing S_k:

Theorem 7 (Recovery of the shared latent). *We have the following lower bound over the weighted factor alignment objective:*

$$\frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k|-1)} \mathcal{L}_{rep,k} \geq \log(L) - \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k|-1)} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} [I(g_{\psi_{i,k}}(O_i); S_k) - I(g_{\psi_{i,k}}(O_i); S_k | g_{\psi_{j,k}}(O_j))] \tag{29}$$

Proof. From Lemma 2, we have:

$$\mathcal{L}_{\text{simCLR}}(g_{\psi_{i,k}}, g_{\psi_{j,k}}) \geq \log(L) - I(Z_i; Z_j) \tag{30}$$

$$\frac{1}{2} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} \Big[\mathcal{L}_{\text{simCLR}}(g_{\psi_{i,k}}, g_{\psi_{j,k}}) + \mathcal{L}_{\text{simCLR}}(g_{\psi_{j,k}}, g_{\psi_{i,k}}) \Big] \geq |\mathcal{I}_k|(|\mathcal{I}_k| - 1) \log(L) - \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} I(Z_{i,k}; Z_{j,k}) \tag{31}$$

$$\frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \mathcal{L}_{\text{rep},k} \geq \log(L) - \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} I(Z_{i,k}; Z_{j,k}) \tag{32}$$

$$\frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \mathcal{L}_{\text{rep},k} \geq \log(L) - \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} [I(g_{\psi_{i,k}}(O_i); S_k) - I(g_{\psi_{i,k}}(O_i); S_k | g_{\psi_{j,k}}(O_j))] \tag{33}$$

Summing Theorem 7 over all shared latents $k \in \{1, \ldots, K\}$ yields a global bound on the full CoACT objective:

Corollary 1. *The CoACT loss satisfy the following:*

$$\mathcal{L}_{CoACT} \geq K \log(L) - \sum_{\substack{k \in \{1,\ldots,K\} \\ |\mathcal{I}_k| \geq 2}} \frac{1}{|\mathcal{I}_k|(|\mathcal{I}_k| - 1)} \sum_{\substack{i,j \in \mathcal{I}_k^2 \\ i \neq j}} [I(Z_{i,k}; S_k) - I(Z_{i,k}; S_k | Z_{j,k})] \tag{34}$$

Remark 4. Assumption 4 requires that agents in $\mathcal{I}_k$ observe S_k through independent noise processes, which is a reasonable model for spatially separated agents with local observations. A necessary condition for this to hold across all factors is that any pair of agents shares at most one latent factor, i.e. $|\mathcal{I}_k \cap \mathcal{I}_{k'}| \leq 1$ for all $k \neq k'$. When this is violated, the representations may capture spurious cross-factor correlations, leading to only partial recovery of the latent factors. Similarly to Remark 1, the hyperparameter α can be decayed to prioritize $\mathcal{L}_{\text{RL}}$ at later stages of training, mitigating the impact of such violations.

Bibliography

[1] Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al.: Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680 (2019)

[2] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)

[3] Chevalier-Boisvert, M., Dai, B., Towers, M., de Lazcano, R., Willems, L., Lahlou, S., Pal, S., Castro, P.S., Terry, J.: Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. CoRR **abs/2306.13831** (2023)

[4] Chu, T., Wang, J., Codecà, L., Li, Z.: Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE transactions on intelligent transportation systems **21**(3), 1086–1095 (2019)

[5] Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012). https://doi.org/10.1109/MSP.2012.2211477

[6] Federici, M., Dutta, A., Forré, P., Kushman, N., Akata, Z.: Learning robust representations via multi-view information bottleneck. arXiv preprint arXiv:2002.07017 (2020)

[7] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

[8] Gronauer, S., Diepold, K.: Multi-agent deep reinforcement learning: a survey. Artif. Intell. Rev. **55**(2), 895–943 (Feb 2022). https://doi.org/10.1007/s10462-021-09996-w, https://doi.org/10.1007/s10462-021-09996-w

[9] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[10] Hu, J., Jiang, Y., Weng, P.: Revisiting data augmentation in deep reinforcement learning. arXiv preprint arXiv:2402.12181 (2024)

[11] Hüttenrauch, M., Šošić, A., Neumann, G.: Deep reinforcement learning for swarm systems. Journal of Machine Learning Research **20**(54), 1–31 (2019)

[12] Laskin, M., Srinivas, A., Abbeel, P.: Curl: Contrastive unsupervised representations for reinforcement learning. In: International conference on machine learning. pp. 5639–5650. PMLR (2020)

[13] Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P., Srinivas, A.: Reinforcement learning with augmented data. Advances in neural information processing systems **33**, 19884–19895 (2020)

[14] Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Pieter Abbeel, O., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems **30** (2017)

[15] McClellan, J., Haghani, N., Winder, J., Huang, F., Tokekar, P.: Boosting sample efficiency and generalization in multi-agent reinforcement learning via equivariance. Advances in Neural Information Processing Systems **37**, 41132–41156 (2024)
[16] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)
[17] Mordatch, I., Abbeel, P.: Emergence of grounded compositional language in multi-agent populations. arXiv preprint arXiv:1703.04908 (2017)
[18] Oliehoek, F.A., Amato, C.: A concise introduction to decentralized POMDPs. Springer (2016)
[19] van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR **abs/1807.03748** (2018), http://arxiv.org/abs/1807.03748
[20] van der Pol, E., van Hoof, H., Oliehoek, F.A., Welling, M.: Multi-agent mdp homomorphic networks. arXiv preprint arXiv:2110.04495 (2021)
[21] Rashid, T., Samvelyan, M., De Witt, C.S., Farquhar, G., Foerster, J., Whiteson, S.: Monotonic value function factorisation for deep multi-agent reinforcement learning. Journal of Machine Learning Research **21**(178), 1–51 (2020)
[22] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. nature **550**(7676), 354–359 (2017)
[23] Song, H., Feng, M., Zhou, W., Li, H.: Ma2cl: Masked attentive contrastive learning for multi-agent reinforcement learning. arXiv preprint arXiv:2306.02006 (2023)
[24] Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K., Graepel, T.: Value-decomposition networks for cooperative multi-agent learning (2017), https://arxiv.org/abs/1706.05296
[25] Tan, M., et al.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the tenth international conference on machine learning. pp. 330–337 (1993)
[26] Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
[27] Van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)
[28] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., Silver, D.: Grandmaster level in StarCraft II using

multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019). https://doi.org/10.1038/s41586-019-1724-z

[29] Wiering, M.: Multi-agent reinforcement leraning for traffic light control. In: Proceedings of the Seventeenth International Conference on Machine Learning. p. 1151–1158. ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2000)

[30] Xu, Z., Zhang, B., Li, D., Zhang, Z., Zhou, G., Chen, H., Fan, G.: Consensus learning for cooperative multi-agent reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11726–11734 (2023)

[31] Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., Locatello, F.: Multi-view causal representation learning with partial observability. arXiv preprint arXiv:2311.04056 (2023)

[32] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of PPO in cooperative multi-agent games. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=YVXaxB6L2Pl