Intent-Grounded Compositional Communication through Mutual Information in Multi-Agent Teams

Seth Karten[1*] and Katia Sycara[1]

***Abstract*— Fundamentally, emergent communication is a representation learning problem. Typically, it is phrased as a Lewis game, in which participants signal using observational information. In multi-agent reinforcement learning (MARL) with communication, coordination information (ordinal) is often required in addition to referential info about one's observations. The information bottleneck defines a trade-off between complexity and utility. However, in MARL, the information sent, and the information received defines a different Markov network than defined in the traditional information bottleneck problem. Thus, in this work, we define, study, and show how to approach the InfoLewis problem, which defines a signaling trade-off between sending referential complexity and ordinal task-specific utility. We use information theory to introduce information rich, variational compositional communication to adequately embed referential information and to provide a contrastive objective to ground communication in intent-specific features. We test our novel methodology on referential and ordinal multi-agent tasks.**

I. INTRODUCTION

Emergent communication studies the creation of artificial language. Often phrased as a Lewis game, speakers and listeners learn a set of tokens to communicate complex observations [1]. However, in multi-agent reinforcement learning (MARL), agents suffer from partial observability and non-stationarity [2], which aims to be solved with decentralized learning through communication. In the MARL setup, agents, as speakers and listeners, learn a set of tokens to communicate observations, intentions, coordination or other experiences which help facilitate solving tasks [3]–[5]. Agents learn to communicate effectively through a backpropagation signal from their task performance [6]–[11]. This has been found useful for applications in human-agent teaming [4], [12]–[14], multi-robot navigation [11], and coordination in complex games such as StarCraft II [15].

A meta-analysis of human studies shows that communication quality has a strong relationship with task performance [16]. When determining a message representation, encoding solely the observations results in suboptimal performance [9]. Recurrent agents are able to process coordination information into their messages, but often send null messages, creating degenerate communication protocols [3], [4], [10]. In an aim to increase the informativeness of communication, recent work has attempted to increase the representational capacity by decreasing the convergence rates [3], [17]–[20]. However, these methods only account for representing *observations* more informatively.

[1] The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. skarten@cs.cmu.edu

Traditionally, in MARL with communication, the communication system is learned in an unsupervised manner from a gradient signal based on the actions taken for the task. However, choosing the correct action relies on a sufficient communication protocol, creating non-stationarity. In this work, we aim to ground the communication to more accurately represent the intent through goal-grounded contrastive learning. Contrastive learning [21], which builds on the MaxEnt reinforcement learning objective [22], aims to build current representations which are closer to future states than random states. We introduce compositional emergent communication grounded in task specific information through contrastive learning.

Mutual information, denoted as $I(X;Y)$, looks to measure the relationship between random variables,

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)}{p(x)}\right] = \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right]$$

which is often measured through Kullback-Leibler divergence [23], $I(X;Y) = D_{KL}(p(x,y)||p(x) \otimes p(y))$. Ultimately, the input message similarity and goal-grounded information can be modeled as the information bottleneck [24], which defines a trade-off between complexity of information (compression, $I(X,\hat{X})$) and the preserved relevant information (utility, $I(\hat{X},Y)$). We introduce an information theoretic objective for multi-agent communication, which shows that one only needs decentralized training for referential, observation-based communication, but centralized training is necessary to properly learn non-referential, emergent communication with respect to coordination and intent. That is, backpropagating the gradient signal through communication edges is essential centralization during training.

In addition, the representational capacity of a token depends on its form and properties. Inspired by continuous word embeddings in natural language, VQ-VIB [20] learns to cluster continuous representations into discrete categories while measuring similar levels of informativeness as continuous tokens. Compositional language strings together multiple tokens to form a single message at each time-step [25] rather than single tokens at each time-step [26]. Compositional language has been shown to promote few-shot generalization to new concepts with humans [13] and agents [27].

In this work, we enable a compositional emergent communication paradigm, which exhibits clustering and informativeness properties. We show theoretically and through empirical results that compositional language enables independence properties among tokens with respect to referential information. Additionally, when combined with contrastive

learning, our method outperforms competing methods that only ground communication on referential information. Finally, we show that contrastive learning acts as an optimal critic for communication, reducing sample complexity for the unsupervised emergent communication objective.

II. RELATED WORK

A. Emergent Communication

Several methodologies currently exist to increase the informativeness of emergent communication. With discrete and clustered continuous communication, the number of observed distinct communication tokens is far below the number permissible [28]. As an attempt to increase the emergent "vocabulary", as well as decrease the data required to converge to an informative communication "language", work has added a bias loss to emit distinct tokens in different situations [17]. Though more recent work has found that the sample efficiency can be further improved by ground communication in observation space with a supervised reconstruction loss [18]. Information maximizing autoencoders aim to maximize the state reconstruction accuracy for each agent. However, grounding communication in observations has found to easily satisfy this objective while still requiring a myriad more samples to explore to find a task-specific communication space [3]. Thus, it is necessary to use task specific information to communicate informatively. Other work aims to use the information bottleneck [24] to decrease the entropy of messages [19]. In our work, we use contrastive learning to increase representation similarity with future goals, which we show optimally optimizes the Q-function for messages.

B. Natural Language Inspiration

The properties of the tokens in emergent communication directly affect their informative ability. As a baseline, continuous communication tokens can represent maximum information but lack human-interpretable properties. Discrete 1-hot (binary vector) tokens allow for a finite vocabulary, but each token contains the same magnitude of information, with equal orthogonal distance to each other token. Similar to word embeddings in natural language, discrete prototypes are an effort to cluster similar information together from continuous vectors [28]. Building on the continuous word embedding properties, VQ-VIB [20], an information theoretic, observation grounding based on VQ-VAE properties [29], uses variational properties to provide word embedding properties for continuous emergent tokens. Like discrete prototypes, they exhibit a clustering property based on similar information, but are more informative. However, each of these message types determine a single token for communication. In our work, we build off VQ-VAE for use in creating compositional communication. Tokens are stringed together to create emergent "sentences".

III. PRELIMINARIES

We formulate our setup as a centralized training, decentralized execution [6], partially observable Markov Decision Process with communication (Dec-POMDP-Comm). Formally, our problem is defined by the tuple, $\langle \mathcal{S}, \mathcal{A}, \mathcal{M}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \Omega, \gamma \rangle$. We define $\mathcal{S}$ as the set of states, $\mathcal{A}^i$, $i \in [1, N]$ as the set of actions, which includes task specific actions, and $\mathcal{M}^i$ as the set of communications for N agents. $\mathcal{T}$ is the transition between states due to the multi-agent joint action space $\mathcal{T} : \mathcal{S} \times \mathcal{A}^1, ..., \mathcal{A}^N \rightarrow \mathcal{S}$. Ω defines the set of observations in our partially observable setting. The partial observability requires communication to complete the tasks successfully. $\mathcal{O}^i : \mathcal{M}^1, ..., \mathcal{M}^N \times \hat{\mathcal{S}} \rightarrow \Omega$ maps the communications and local state, $\hat{\mathcal{S}}$, to a distribution of observations for each agent. $\mathcal{R}$ defines the reward function and γ defines the discount factor.

A. Architecture

We build on REINFORCE [30]. However, since we study cooperative tasks, we allow agents to share their policy network parameters during training. The policy network is defined by three stages: Observation encoding, Communication, and action decoding. The best observation encoding and action decoding architecture is task dependent, i.e., using multi-layer perceptrons (MLPs), CNNs [31], GRUs [32], or transformer [33] layers are best suited to different inputs. The encoder transforms observation and any sequence or memory information into an encoding H.

Our work focuses on the communication stage, which can be into three substages: message encoding, message passing (often considered sparse communication), and message decoding. The message passing is defined by concurrent work [34]. For message decoding, we build on a multi-headed attention framework, which allows an agent to learn which messages are most important [35]. The message encoding is defined by our compositional communication framework, as described in section IV.

B. Objective

The message encoding substage can be defined as an information bottleneck problem. The deep variational information bottleneck defines a trade-off between preserving useful information and compression [24], [36]. We assume that our observation and memory/sequence encoder provides an optimal representation H^i suitable for sharing relevant observation and intent/coordination information. We hope to recover a representation Y^i, which contains the sufficient desired outputs.

In our scenario, the information bottleneck is a trade-off between complexity of information $I(H^i; M^i)$ (representing the encoded information exactly) and representing the relevant information $I(M^{j \neq i}; Y^i)$, which is signaled from our contrastive objective. In our setup, the relevant information flows from other agents through communication, signaling a combination information bottleneck and Lewis game, dubbed InfoLewis. We additionally promote complexity through our compositional independence objective, $I(M^i_1; \ldots; M^i_L | H^i)$.

This is formulated by the following Lagrangian,

$$\mathcal{L}(\ p(m^i|h^i)\ ) = \ -\beta_u \hat{I}(M^{j\neq i}; Y^i)\ + \beta_c \hat{I}(H^i; M^i) \\ - \beta_I \hat{I}(M^i_1; \ldots; M^i_L | H^i)$$

where the bounds on mutual information $\hat{I}$ are defined in equations 2 3 6. The first two terms define the InfoLewis objective. Overall, our objective is,

$$J(\theta) = \max_{\pi} \mathbb{E}\left[\sum_{t\in T}\sum_{i\in N} \gamma_t \mathcal{R}(s_t, a_t) + \mathcal{L}(\ p(m_t|h_t)\ )\right] \\ \text{s.t.}(a_t, m_t, h_t) \sim \pi^i, s_t \sim \mathcal{T}(s_{t-1})$$

IV. Complexity through Compositional Communication

We aim to satisfy the complexity objective, $I(H^i, M^i)$, through compositional communication. In order to induce complexity in our communication, we want the messages to be as non-random as possible. That is, informative with respect to the input hidden state h. In addition, we want each token within the message to share as little information as possible with the preceding tokens. Thus, each additional token adds *only informative* content. Each token has a fixed length in bits W. The total sequence is limited by a fixed limit, $\sum_l^L W_l \leq S$, of S bits and a total of L tokens.

We use a variational message generation setup based on VQ-VAE [29], which maps the encoded hidden state h to a message m; that is, we are modeling the posterior, $\pi^i_m(m_l|h)$. We limit the vocabulary size to K tokens, $e_j \in \mathbb{R}^D, j \in [1, K] \subset \mathbb{N}$, where each token has dimensionality D and $l \in [1, L] \subset \mathbb{N}$. Each token m_l is sampled from a categorical posterior distribution,

$$\pi^i_m(m_l = e_k|h) = \begin{cases} 1 & \text{for } k = \arg\min_j ||m_l - e_j||_2 \\ 0 & \text{otherwise} \end{cases}$$

such that the message m_l is mapped to the nearest neighbor e_j. A set of these tokens makes a message m. To satisfy the complexity objective, we want to use m^i to well-represent h^i and consist of independently informative m^i_l.

A. Independent Information

Starting with the independent information objective, we want to minimize the interaction information,

$$I(m_1; \ldots; m_L|h) = \\ \int \ldots \int f_m(m_1, \ldots, m_L, h) dh\ dm_1 \ldots dm_L$$

which defines the conditional mutual information between each token and,

$$f_m(*) = p(h)p(m_1; \ldots; m_L|h) \log \frac{p(m_1; \ldots; m_L|h)}{\prod_l^L p(m_L|h)} \quad (1)$$

Let $\pi^i_m(m_l|h)$ be a variational approximation of $p(m_l|h)$, which is defined by our message encoder network. Given that each token should provide unique information, we assume independence between c_m. Thus, it follows that our compositional message is a vector, $m = [m_1, \ldots, m_L]$, and is jointly Gaussian. Moreover, we can define $q(\hat{m}|h)$ as a variational approximation to $p(m|h) = p(m_1; \ldots, m_L|h)$. We can model q with a network layer and define its loss as $||\hat{m} - m||_2$. Thus, transforming equation 1 into variational form, we have,

$$g_m(m_1, \ldots, m_L, h) = p(h)q(\hat{m}|h) \log \frac{q(\hat{m}|h)}{\prod_l^L \pi^i_m(m_l|h)}$$

Since Kullback Leibler divergence D_{KL} is non-negative, $D_{KL}\left(q(\hat{m}|h)||\pi^i_m(m_1|h) \otimes \cdots \otimes \pi^i_m(m_L|h)\right) \geq 0$, it follows that $\int q(\hat{m}|h) \log q(\hat{m}|h) d\hat{m} \geq \int q(\hat{m}|h) \log \prod_l^L \pi^i_m(m_l|h) d\hat{m}$ Thus, we can lower bound our interaction information,

$$I(m_1; \ldots; m_L|h) \geq \int \ldots \int g_m(*) dh dm_1 \ldots dm_L \\ = \mathbb{E}_{h\sim p(h)}\left[D_{KL}\left(q(\hat{m}|h)||\pi^i_m(m_1|h) \otimes \cdots \otimes \pi^i_m(m_L|h)\right)\right]$$

Since we want the mutual information to be minimized in our objective, we maximize,

$$\hat{I}(m_1; \ldots; m_L|h) = \\ \mathbb{E}_{h\sim p(h)}\left[D_{KL}\left(q(\hat{m}|h)||\pi^i_m(m_1|h) \otimes \cdots \otimes \pi^i_m(m_L|h)\right)\right] \quad (2)$$

B. Input-Oriented Information

In order to induce complexity in the compositional messages, we additionally want to maximize the mutual information $I(H; M)$ between the composed message $\hat{m}$ and the encoded information h. By definition of mutual information, we have,

$$I(H; M) = \int \int p(h)p(\hat{m}|h) \log \frac{p(\hat{m}|h)}{p(\hat{m})} d\hat{m}\ dh$$

Substituting $q(\hat{m}|h)$ for $p(\hat{m}|h)$, the same KL Divergence identity, and defining a Gaussian approximation $z(\hat{m})$ of the marginal distribution $p(\hat{m})$, it follows that,

$$I(H; M) \geq \int \int p(h)q(\hat{m}|h) \log \frac{q(\hat{m}|h)}{z(\hat{m})} d\hat{m}\ dh$$

In expectation of equation 2, we have $q(\hat{m}|h) = q(\hat{m}|h) = \prod_l^L \pi^i_m(m_l|h)$. This implies that, for $\hat{m} = [m_1, \ldots, m_L]$, there is probabilistic independence between $m_j, m_k, j \neq k$. Thus, expanding, it follows that,

$$I(H; M) \geq \sum_l^L \int \int p(h)q(m_l|h) \log \frac{q(m_l|h)}{z(m_l)} dm_l\ dh \\ = \sum_l^L \mathbb{E}_{h\sim p(h)}\left[D_{KL}\left(q(m_l|h)||z(m_l)\right)\right]$$

where $z(m_l)$ is a standard Gaussian. Thus, we have our Langrangian term,

$$\hat{I}(H^i, M^i) = -\sum_l^L \mathbb{E}_{h\sim p(h)}\left[D_{KL}\left(q(m_l|h)||z(m_l)\right)\right] \quad (3)$$

Conditioning on the input or observation data is a decentralized training objective.

C. Message Generation Architecture

Now, we can define the pipeline for message generation. The idea is to create an architecture that can generate features to enable independent message tokens. We expand each compressed token into the space of the hidden state h (1-layer linear expansion) since each token has a natural embedding in $\mathbf{R}^{|h|}$. Then, we perform attention using a `softmin` to help minimize similarity with previous tokens and sample the new token from a variational distribution. See algorithm 1 for full details. During execution, we can generate messages directly due to equation 2, resolving any computation time lost from sequential compositional message generation.

Algorithm 1 `Compositional Message Gen.`(h_t)

$T \leftarrow$ `num_tokens`
$m = \mathbf{0}$ ▷ $T \times d_m$, $d_m \leftarrow$ `token_size`
$Q \leftarrow$ `Q_MLP`(h_t)
$V \leftarrow$ `V_MLP`(h_t)
for $i \leftarrow 1$ to T **do**
  $K \leftarrow$ `K_MLP`(m)
  $\hat{h} = \texttt{softmin}(\frac{Q^\intercal \texttt{mean}(K,1)}{\sqrt{d_k}})^\intercal V$
  $m_i \sim \mathcal{N}(\hat{h}; \mu, \sigma)$
end for
return m

V. Utility through Contrastive Learning

First, note that our Markov Network is as follows: $H^j \rightarrow M^j \rightarrow Y^i \leftarrow H^i$. Continue to denote i as the agent identification and j as agent ID such that $j \neq i$. We aim to satisfy the utility objective of the information bottleneck, $I(M^j; Y^i)$, through contrastive learning.

Proposition V.1. *Utility mutual information is lower bounded by the contrastive NCE-binary objective,* $I(M, Y) \geq \log \sigma(f(s, m, s_f^+)) - \log \sigma(f(s, m, s_f^-))$.

We suppress the reliance on h since this is directly passed through. By definition of mutual information, we have,

$$I(M^j; Y^i) = \int \int p(m) \pi_{R^+}(y|m) \log \frac{\pi_{R^+}(y|m)}{\pi_{R^-}(y)} dm\, dy$$

Our network model learns $\pi_{R^+}(y|m)$ from rolled-out trajectories, R^+, using our policy. The prior of our network state, $\pi_{R^-}(y)$, can be modeled from rolling out a random trajectory, $R-$. Unfortunately, it is intractable to model $\pi_{R^+}(y|m)$ and $\pi_{R^-}(y)$ directly during iterative learning, but we can sample $y^+ \sim \pi_{R^+}(y|m)$ and $y^- \sim \pi_{R^-}(y)$ directly from our network during training.

It has been shown that $\log p(y|m)$ provides a lower bound on mutual information [37],

$$I(M^j; Y^i) \geq \mathbb{E}\left[\frac{1}{K} \sum_{k=1}^{K} \log \pi_{R^+}(y_k|m_k) - \log \pi_{R^-}(y_k)\right] \tag{4}$$

with the expectation over $\prod_l p(m_l, y_l)$. However, we need a tractable understanding of the information Y.

Lemma V.2. $\pi_{R^-}(y) = p(s' = s_f^-|y)$.

In the information bottleneck, Y represents a desired outcome. In our setup, y is coordination information which helps create a desired out, such as any action a^-. This implies, $y \implies a^-$. Since the transition is known, it follows that $a^- \implies s_f^-$, a random future state. Thus, we have, $\pi_{R^-}(y) = p(s' = s_f^-|y)$.

Lemma V.3. $\pi_{R^+}(y|m) = p(s' = s_f^+|y, m)$.

This is similar to the proof for lemma V.2, but requires assumptions on messages m from the emergent language. We note that when m is random, the case defaults to lemma V.2. Thus, we assume we have at least input-oriented information in m given sufficiently satisfying equation 3. Given a sufficient emergent language, it follows that $y \implies a^+$, where a^+ is an intention action based on m. Similarly, since the transition is known, $a^+ \implies s_f^+$, a desired goal state along the trajectory. Thus, we have, $\pi_{R^+}(y|m) = p(s' = s_f^+|y, m)$.

Recall the following (as shown in [21]), which we have adapted to our communication objective,

Proposition V.4 (rewards $\rightarrow$ probabilities)**.** *The Q-function for the goal-conditioned reward function* $r_g(s_t, m_t) = (1 - \gamma)p(s' = s_g|y_t)$ *is equivalent to the probability of state* s_g *under the discounted state occupancy measure:*

$$Q_{s_g}^{\pi}(s, m) = p^{\pi}(s_f^+ = s_g|y) \tag{5}$$

and

Lemma V.5. *The critic function that optimizes equation 4 is a Q-function for the goal-conditioned reward function up to a multiplicative constant* $\frac{1}{p(s_f)}$*:* $\exp(f^*(s, m, s_f) = \frac{1}{p(s_f)} Q_{s_f}^{\pi}(s, m)$.

The critic function $f(s, m, s_f) = y^\intercal \texttt{enc}(s_f)$ represents the similarity between the encoding $y = \texttt{enc}(s, m)$ and the encoding of the future rollout s_f.

Given lemmas V.2 V.3 V.5 and proposition V.4, it follows that equation 4 is the NCE-binary [38] (InfoMAX [39]) objective,

$$\hat{I}(M^j, Y^i) = \log\Big(\sigma(f(s, m, s_f^+))\Big) + \log\Big(1 - \sigma(f(s, m, s_f^-))\Big) \tag{6}$$

which lower bounds the mutual information, $I(M^j, Y^i) \geq \hat{I}(M^j, Y^i)$. The critic function is unbounded, so we constrain it to $[0, 1]$ with the sigmoid function, $\sigma(*)$.

This result shows a need for gradient information to flow backwards across agents along communication edge connections.

VI. Experiments and Results

When evaluating an artificial language in MARL, we only are interested in referential tasks, in which communication is *required* to complete the task. With regard to intent-grounded communication, we study ordinal tasks, which require coordination information between agents to successfully complete. Thus, we consider tasks with a team of agents to foster messaging that communicates coordination information that also includes their observations.

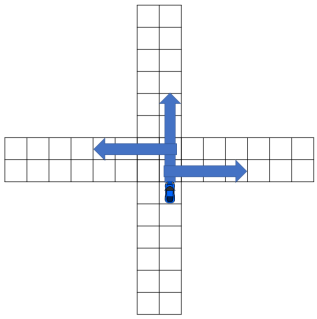

Fig. 1: 10 agents navigate without vision of other agents through the bidirectional traffic junction environment.

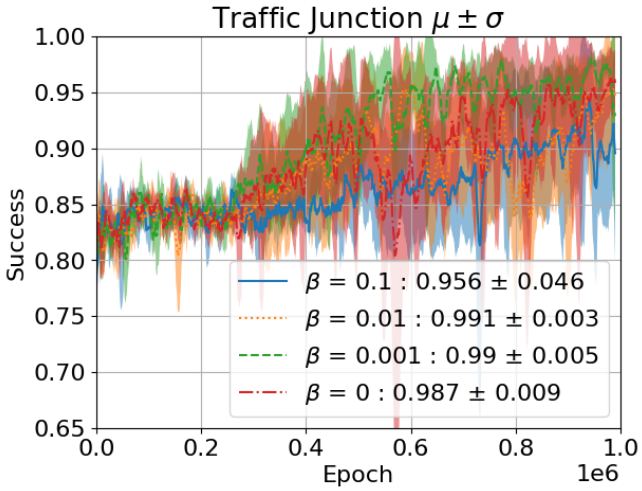

Fig. 2: Above is an ablation of the derived referential complexity loss for our compositional communication. The legend provides mean ± variance of the best performance.

A. Environments

We consider a benchmark which requires both referential and ordinal capabilities within a team of agents. The blind traffic junction environment [10] requires multiple agents to navigate a junction without any observation of other agents. Rather, they only observe their own state location. See figure 1. We evaluate over 10 seeds.

B. Baselines

To evaluate the utility of our contrastive objective, we compare our method to the following baselines: (1) `no-comm`, where agents do not communicate; (2) `rl-comm`, which uses a baseline communication method learned solely through policy loss [10]; (3) `ae-comm`, which uses an autoencoder to ground communication in input observations [18]; (4) `VQ-VIB`, which uses a variational autoencoder to ground discrete communication in input observations and a mutual information objective to ensure low entropy communication [20].

C. Input-Oriented Information Results

We provide an ablation of the loss parameter β in Fig. 2. When $\beta = 0$, we use our compositional message paradigm without our derived loss terms. We find that higher complexity and independence loss increases sample complexity.

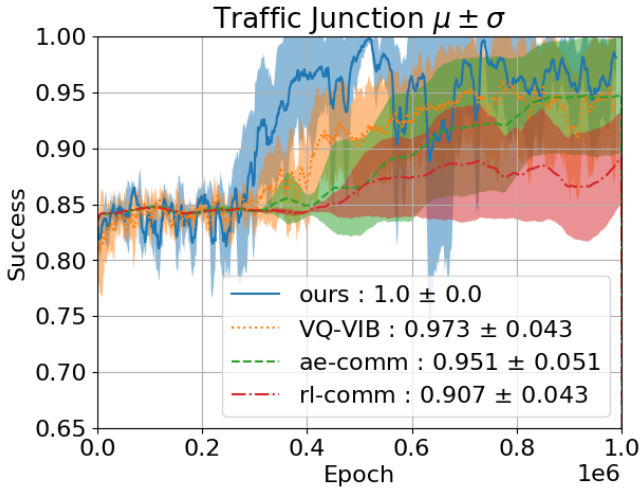

Fig. 3: Our method uses compositional complexity and contrastive utility to outperform other baselines in terms of performance and sample complexity. The legend provides mean ± variance of the best performance.

When $\beta = 1$, the model was unable to converge. However, when there is no regularization loss, the model performs worse (and has no guarantees about referential representation).

D. Communication Utility Results

Unfortunately, due to coordination in MARL, grounding communication in referential features is not enough. Finding the communication utility requires grounding messages in ordinal information. Overall, Fig. 3 shows that our compositional, contrastive method outperforms all methods focused on solely input-oriented communication grounding. Our method yields a higher average task success rate and is able to achieve it with a lower sample complexity. Training with the contrastive update tends to spike to high success but not converge, often many episodes before convergence, which leaves area for training improvement.

a) Regularization loss convergence: At convergence to high task performance, the autoencoder loss actually increases in order to represent the coordination information. This follows directly from the information bottleneck. However, our compositional communication loss does not converge before task performance convergence. Additionally, the contrastive loss tends to monotonically decrease and converges after the task performances converges. This implies empirical evidence that the contrastive loss is an optimal critic for messaging.

VII. Discussion

Any referential-based setup can be performed with a supervised loss, as indicated by the instant satisfaction of referential objectives. However, in multi-agent settings, the harder challenge is to enable coordination through communication. Using contrastive communication as an optimal critic aims to satisfy this. Since contrastive learning benefits from good examples, this method would perhaps be even more powerful in offline RL. In this setting, the communication may be bootstrapped, since our optimal critic has examples

with strong signals. Additionally, the minimization of our independence objective enables tokens which contain minimal overlapping information with other tokens. Preventing trivial communication paradigms enables higher performance. We want to continue to explore the benefits of these properties in future applications. Each of these objectives is complementary, so they are not trivially minimized during training, which is a substantial advantage over comparative baselines. Unlike prior work, this enables the benefits of training with reinforcement learning in multi-agent settings.

REFERENCES

[1] D. Lewis, *Convention*. Cambridge, MA: Harvard University Press, 1969. I

[2] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, "Dealing with non-stationarity in multi-agent deep reinforcement learning," *arXiv preprint arXiv:1906.04737*, 2019. I

[3] S. Karten, M. Tucker, S. Kailas, and K. Sycara, "Towards true lossless sparse communication in multi-agent systems," *preprint*, 2022. I, II-A

[4] S. Karten, M. Tucker, H. Li, S. Kailas, M. Lewis, and K. Sycara, "Interpretable learned emergent communication for human-agent teams," *preprint*, 2022. I

[5] C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent reinforcement learning with communication," *arXiv preprint arXiv:2203.08975*, 2022. I

[6] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2145–2153. I, III

[7] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6382–6393. I

[8] A. Lazaridou, A. Peysakhovich, and M. Baroni, "Multi-agent cooperation and the emergence of (natural) language," *arXiv preprint arXiv:1612.07182*, 2016. I

[9] S. Sukhbaatar, R. Fergus, *et al.*, "Learning multiagent communication with backpropagation," *Advances in neural information processing systems*, vol. 29, pp. 2244–2252, 2016. I

[10] A. Singh, T. Jain, and S. Sukhbaatar, "Learning when to communicate at scale in multiagent cooperative and competitive tasks," in *International Conference on Learning Representations*, 2018. I, VI-A, VI-B

[11] B. Freed, R. James, G. Sartoretti, and H. Choset, "Sparse discrete communication learning for multi-agent cooperation through backpropagation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 7993–7998. I

[12] A. R. Marathe, K. E. Schaefer, A. W. Evans, and J. S. Metcalfe, "Bidirectional communication for effective human-agent teaming," in *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 2018, pp. 338–350. I

[13] B. M. Lake, T. Linzen, and M. Baroni, "Human few-shot learning of compositional instructions," *arXiv preprint arXiv:1901.04587*, 2019. I

[14] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," *arXiv preprint arXiv:2006.02419*, 2020. I

[15] M. Samvelyan, T. Rashid, C. S. De Witt, G. Farquhar, N. Nardelli, T. G. Rudner, C.-M. Hung, P. H. Torr, J. Foerster, and S. Whiteson, "The starcraft multi-agent challenge," *arXiv preprint arXiv:1902.04043*, 2019. I

[16] S. L. Marlow, C. N. Lacerenza, J. Paoletti, C. S. Burke, and E. Salas, "Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance," *Organizational behavior and human decision processes*, vol. 144, pp. 145–170, 2018. I

[17] T. Eccles, Y. Bachrach, G. Lever, A. Lazaridou, and T. Graepel, "Biases for emergent communication in multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 32, 2019. I, II-A

[18] T. Lin, J. Huh, C. Stauffer, S. N. Lim, and P. Isola, "Learning to ground multi-agent communication with autoencoders," *Advances in Neural Information Processing Systems*, vol. 34, 2021. I, II-A, VI-B

[19] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, "Learning efficient multi-agent communication: An information bottleneck approach," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9908–9918. I, II-A

[20] M. Tucker, J. Shah, R. Levy, and N. Zaslavsky, "Towards human-agent communication via the information bottleneck principle," *arXiv preprint arXiv:2207.00088*, 2022. I, II-B, VI-B

[21] B. Eysenbach, T. Zhang, R. Salakhutdinov, and S. Levine, "Contrastive learning as goal-conditioned reinforcement learning," *arXiv preprint arXiv:2206.07568*, 2022. I, V

[22] B. Eysenbach and S. Levine, "Maximum entropy rl (provably) solves some robust rl problems," in *International Conference on Learning Representations*, 2021. I

[23] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997. I

[24] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*. IEEE, 2015, pp. 1–5. I, II-A, III-B

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013. I

[26] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018. I

[27] Ł. Kuciński, T. Korbak, P. Kołodziej, and P. Miłoś, "Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 075–23 088, 2021. I

[28] M. Tucker, H. Li, S. Agrawal, D. Hughes, K. Sycara, M. Lewis, and J. A. Shah, "Emergent discrete communication in semantic spaces," *Advances in Neural Information Processing Systems*, vol. 34, 2021. II-A, II-B

[29] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017. II-B, IV

[30] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992. III-A

[31] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995. III-A

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014. III-A

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. III-A

[34] S. Karten, D. Hughes, and K. Sycara, "Who, what, when? an overarching guide to sparse communication in multi-agent teams," *preprint*, 2022. III-A

[35] A. Agarwal, S. Kumar, K. Sycara, and M. Lewis, "Learning transferable cooperative behavior in multi-agent teams," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020, pp. 1741–1743. III-A

[36] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *ICLR*, 2017. III-B

[37] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5171–5180. V

[38] Z. Ma and M. Collins, "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," in *EMNLP*, 2018. V

[39] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2018. V