
Semantics and Spatiality of Emergent Communication

Rotem Ben Zion¹ **Boaz Carmeli¹** **Orr Paradise²** **Yonatan Belinkov¹**

¹ Technion – Israel Institute of Technology ² UC Berkeley

`rotm@campus.technion.ac.il` `boaz.carmeli@campus.technion.ac.il`
`orrp@eecs.berkeley.edu` `belinkov@technion.ac.il`

Abstract

When artificial agents are jointly trained to perform collaborative tasks using a communication channel, they develop opaque goal-oriented communication protocols. Good task performance is often considered sufficient evidence that meaningful communication is taking place, but existing empirical results show that communication strategies induced by common objectives can be counterintuitive whilst solving the task nearly perfectly. In this work, we identify a goal-agnostic prerequisite to meaningful communication, which we term *semantic consistency*, based on the idea that messages should have similar meanings across instances. We provide a formal definition for this idea, and use it to compare the two most common objectives in the field of emergent communication: discrimination and reconstruction. We prove, under mild assumptions, that semantically inconsistent communication protocols can be optimal solutions to the discrimination task, but not to reconstruction. We further show that the reconstruction objective encourages a stricter property, *spatial meaningfulness*, which also accounts for the distance between messages. Experiments with emergent communication games validate our theoretical results. These findings demonstrate an inherent advantage of distance-based communication goals, and contextualize previous empirical discoveries.

1 Introduction

Humans use language in multi-agent social interactions. Pressures of synchronization and collaboration play a central role in shaping the way we communicate. Motivated by this observation and the goal of creating artificial agents capable of meaningful communication, the field of emergent communication (EC) employs a multi-agent environment jointly trained to accomplish a task that requires active transmission of information. The agents utilize a messaging channel that usually takes the form of a discrete sequence of abstract symbols, resembling the structure of human language. Successful optimization results in synchronized agents operating a newly developed communication protocol tailored to the objective. We study a type of EC setup inspired by Lewis games [32] where a sender agent describes a given input and a receiver agent makes a prediction based on that description. The game objective is designed to make the receiver demonstrate knowledge of the original input, which in turn compels the sender to encode relevant information in the message. There are two common objectives used in this type of EC setup (illustrated in Figure 1):

Reconstruction In the reconstruction task [34, 44, 45], the original input is re-generated based solely on the message. We are specifically interested in a reconstruction objective that quantifies distance between prediction and target, forming a discrete version of autoencoding [18].

Discrimination In the discrimination task [17, 28, 33], the original input is retrieved from a set of candidates, incentivized by negative log-likelihood.

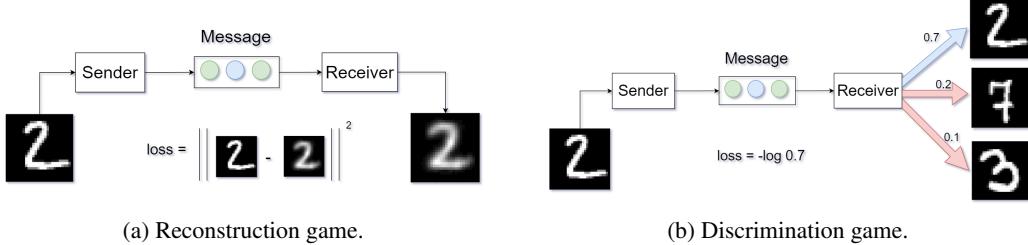


Figure 1: Illustration of the reconstruction and discrimination tasks. The discrimination receiver is given the candidates in addition to the message.

A central goal in EC, which motivates our work, is understanding how different factors in the environment affect the emergent protocol, and specifically developing agents and objectives that create protocols with similar characteristics to natural language [2, 25, 40, 41, 45]. Tools and experiments have been developed to evaluate the proximity to natural language by testing for properties such as compositionality [6] or efficiency [7]. Unfortunately, many of these empirical methods show great dissimilarity to human communication. One particularly surprising experiment [3] revealed that protocols created via the discrimination task can generalize extremely well to random noise data, suggesting that they do not signal human-recognizable (meaningful) properties of the input.

In this paper, we identify a fundamental property of any meaningful communication protocol, and thus a prerequisite for similarity to natural language. We observe that the discrete bottleneck forces EC protocols to be many-to-one mappings, i.e., that messages likely represent more than one input. With this in mind, we claim that inputs mapped to the same message should be *semantically similar*, as is the case with human language. We formalize this idea with a mathematical definition that we term **semantic consistency**. We further develop a stricter version of this definition, **spatial meaningfulness**, which also accounts for distances between messages, and is therefore better suited to indicate similarity to natural language.

Armed with these definitions, we analyze theoretical solutions to the two common EC environments. In the reconstruction setting, under mild assumptions, we prove that **every optimal solution is semantically consistent**. With a different set of assumptions, we also show that reconstruction-induced communication protocols are spatially meaningful. In sharp contrast, we surprisingly find that the discrimination objective **does not guarantee semantic consistency**, i.e., a communication protocol can be optimal in a discrimination environment but still not semantically consistent nor spatially meaningful. In fact, we show that uniformly random messages can lead to a globally optimal discrimination objective value, meaning that the relationship between inputs mapped to the same message is potentially arbitrary despite optimally solving the task. Our results provide theoretical support to previous empirical findings, such as the discrimination generalization to noise. We further analyze several common variations of the discrimination game from the EC literature.

To support our findings, we run experiments on Shapes [27] and MNIST [30]. The empirical results agree with most of our theoretical findings. However, we observe a gap between theory and practice regarding the level of channel utilization, which we leave for future work to further investigate.

2 Background and related work

2.1 Emergent communication with Lewis games

A large portion of EC research, based on Lewis signaling games [28, 32], defines two-agent cooperative tasks where one agent, S_θ (“sender”), receives an input $x \in \mathcal{X}$ (“target”) and generates a message $m \in M$, and the second agent, R_φ (“receiver”), takes that message and outputs a prediction. The two most common such tasks are reconstruction and discrimination.

In the reconstruction task, the receiver tries to predict the target directly from the message. This is a generative method, so it can naturally be modeled with a receiver that outputs a distribution over the input space [7, 41, 42, 44], i.e., $R_\varphi : M \rightarrow \Delta(\mathcal{X})$. In this case, the objective is usually negative log-likelihood or accuracy. However, this approach has two major flaws: First, explicitly outputting a distribution is not always feasible, e.g., over images. Thus, this game is mostly implemented on simple synthetic datasets. Second, the incentive of likelihood or accuracy does not measure *how*

close the receiver’s output is to x .¹ In this work, we assume a different approach to reconstruction, inspired by autoencoders [18, 34], where the receiver outputs an object in the input space rather than a distribution, i.e., $R_\varphi : M \rightarrow \mathcal{X}$. Thus, the loss function in our analysis is straightforwardly the distance between prediction and target.

In the discrimination task, the receiver tries to identify the target within a set of candidates [17, 28, 29, 33, 44]. It can be modeled as $R_\varphi : M \times \mathcal{X}^d \rightarrow \Delta\{1, \dots, d\}$. We explore some alternative formulations in Appendix F. This setting introduces two major choices: the number of distractors (candidates other than x) and how to select them [12, 29]. These parameters have been shown to have a major effect on the resulting protocols [15, 29]. The results in the main body of this paper concern the vanilla version of the discrimination task, where candidates are chosen independently. We analyze other discrimination versions in Appendix A.

2.2 Towards interpretable human-like communication

A central line of study aims to develop EC systems that create protocols with high proximity to natural language. Some have tried aligning the two explicitly, by using pretrained language models [31, 43] or training translation systems [45, 46]. In this work, we focus on properties of the EC protocol itself, as induced by the objective rather than exposure to text. The most studied such characteristic is compositionality, which refers to the ability of agents to assemble complex units (e.g., sentences) from simpler units (e.g., words). Compositionality is hard to measure [1, 5, 6, 24, 37], and while some claim that it is correlated with generalization [2, 40, 42], the opposite has also been argued [8, 11, 13]. In any case, emergent languages are often not compositional [25], inefficient [7, 41], and generalize to noise [3], suggesting a mismatch between common EC objectives and the evolution of human language. In this work, we advance the goal of creating human-like communication by interpreting properties of natural language into formal constraints, and analyzing whether the EC objectives create protocols that follow them. A commonly used compositionality measure, topographic similarity (topsim) [5], is related to the definitions presented in this work. It evaluates the correlation between distances in the input space and the corresponding distances in the messages space. Notably, topsim considers the relationship between every pair of inputs, whereas our definitions only consider pairs that correspond to similar messages. The latter follows an intuitive asymmetry: inputs with similar messages are expected to be similar, but inputs with dissimilar messages do not have to be different (for example, the same image can often be described by several distinct human captions).

2.3 Input information encoded in the message

The game objective dictates the information that needs to be stored in messages. In this work, we focus on spatial information by considering the geometric relationship between inputs mapped to the same message. To the best of our knowledge, no previous work has theoretically analyzed this idea in EC literature. A related concept is the mutual information (MI) between messages and inputs, which has been studied both empirically [1, 15, 23, 35, 45] and theoretically [19, 21, 42]. MI is strongly related to the infoNCE objective [38] often used in implementations of the discrimination game. MI can be used to measure the complexity of the communication channel [44, 45], and is correlated with task performance [23, 45] and even compositionality [1]. Finally, The mutual information gap [9] can be used to measure disentanglement [8, 13], which is also related to compositionality. However, Guo et al. [15] show that MI is not sufficient to indicate transfer-learning generalization, and connect that to pairwise distances of inputs mapped to the same message, effectively testing for semantic consistency. The first part of this work, combined with their findings, would suggest that languages developed by agents in the reconstruction setting should be more expressive and generalize better to different tasks.

On the theoretical side, a relevant work is Rita et al. [42], where the global discrimination objective is decomposed into two interpretable components, a co-adaptation loss and MI. The latter is equivalent to our Lemma A.1. In this work, we use such equivalent objectives for both the reconstruction and discrimination tasks to investigate the behaviour of optimal solutions, and specifically whether the tasks encourage semantic consistency.

¹We refer to this game as “global discrimination” and analyze it in Appendix A.1. When the input space is finite, it is equivalent to a special case of the discrimination game.

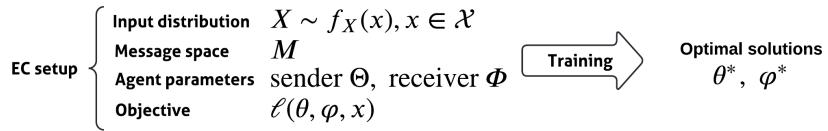


Figure 2: Notation for the emergent communication (EC) setup.

3 Notation and task definitions

3.1 General notation

We begin by introducing general notation for the EC setup; see Figure 2 for illustration. The input space is modeled by a real-valued random variable $X = (\mathcal{X}, f_X)$, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is the bounded set of possible inputs (e.g., images), and $f_X: \mathcal{X} \rightarrow \mathbb{R}_+$ is a prior distribution. Denote by $M = \{m_1, m_2, \dots\} \subset \mathbb{R}^{d_M}$ the finite or countably infinite set of possible messages.² We assume that the minimal distance between a pair of messages is attained, i.e., that $\min_{m_1 \neq m_2 \in M} \|m_1 - m_2\| \in (0, \infty)$.

The *sender* agent $S_\theta: \mathcal{X} \rightarrow M$, parameterized by $\theta \in \Theta$, maps each input to a message. Sender S_θ defines the communication protocol; we compare EC setups based on which optimal sender agents they induce. The *receiver* agent R_φ , maps the message and optionally additional input to a prediction. It is parameterized by $\varphi \in \Phi$. During EC training, the goal is to learn hypotheses (θ, φ) that minimize a *loss* described by a function $\ell: \Theta \times \Phi \times \mathcal{X} \rightarrow \mathbb{R}$, which maps a pair of agents and an input to a real value. To group all of the above notation into a single definition, denote an *EC setup* as a sextuple $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$. An EC setup induces a set of *optimal* agents by

$$\theta^*, \varphi^* \in \arg \min_{\theta \in \Theta, \varphi \in \Phi} \mathbb{E}_{x \sim X} \ell(\theta, \varphi, x).$$

Note that every part of the setup can affect the set of optimal agents. We also define a notion of conditional optimality, where one of the agents is fixed at a potentially sub-optimal hypothesis; sender S_θ is *synchronized* with receiver R_φ if the pair is optimal for the EC setup $(\mathcal{X}, f_X, M, \Theta, \{\varphi\}, \ell)$. The same definition applies to the receiver agent respectively. Note that a pair of agents can be synchronized in both directions without being optimal.

3.2 Task definitions

In the reconstruction setting, the receiver agent predicts the original input (the “target”) given only the message. In the discrimination setting, the receiver is given the message and a set of candidate inputs containing the target in a random position, and attempts to predict that position based on the message. A visual illustration can be seen in Figure 1. We now present formal definitions of these setups.

Reconstruction. The receiver maps a message $S_\theta(x)$ to a prediction in the input space. The loss is the Euclidean distance between that prediction and the target x .

- $\Phi_{\text{reconstruction}}^{M, \mathcal{X}}$ is defined as the set of all functions of the form $R: M \rightarrow \mathcal{X}$.
- $\ell_{\text{reconstruction}}(\theta, \varphi, x) := \|R_\varphi(S_\theta(x)) - x\|^2$.

Definition 1. An EC setup $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ is a *reconstruction game* if $\Phi \subseteq \Phi_{\text{reconstruction}}^{M, \mathcal{X}}$ and $\ell = \ell_{\text{reconstruction}}$. It has an *unrestricted* receiver hypothesis class if $\Phi = \Phi_{\text{reconstruction}}^{M, \mathcal{X}}$.

Discrimination. In addition to a message $S_\theta(x)$, the receiver sees a set of candidates $\{x_1, \dots, x_d\}$, which contains the target at a random position t , i.e. $x_t = x$, and the rest are $d - 1$ independently sampled distractors. The receiver outputs a probability distribution over the candidates. The loss is the negative log-likelihood of the correct position t according to this distribution, averaged over the target position and distractors.

²The description of M as a vector space is flexible. As an example instantiation, the classic communication channel where each message is a sequence of symbols from a fixed vocabulary can be interpreted as a concatenation of one-hot vectors. In that case, we would have $d_M = V \cdot L$ where V is the vocabulary size and L the message length. The Euclidean distance between messages becomes their Hamming distance.

- $\Phi_{\text{discrimination}}^{M, \mathcal{X}, d}$ is defined as the set of all functions of the form $R: M \times \mathcal{X}^d \rightarrow \Delta\{1, \dots, d\}$.
- $\ell_{\text{discrimination}}^{X, d}(\theta, \varphi, x) := \mathbb{E}_{\substack{t \sim U\{1, \dots, d\}, x_t=x \\ x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d \sim X^{d-1}}} - \log P(R_\varphi(S_\theta(x), x_1, \dots, x_d) = t)$.

Definition 2. An EC setup $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ is a d -candidates discrimination game if $\Phi \subseteq \Phi_{\text{discrimination}}^{M, \mathcal{X}, d}$ and $\ell = \ell_{\text{discrimination}}^{X, d}$. It has an unrestricted receiver hypothesis class if $\Phi = \Phi_{\text{discrimination}}^{M, \mathcal{X}, d}$

In this vanilla version of the discrimination game, the candidates are sampled independently. Several other choice mechanisms have been explored in EC literature. In Appendix A, we define and analyze the most common such variations:

Global discrimination In Appendix A.1, we analyze a version of the discrimination game where the receiver outputs a distribution over the entire data rather than a small set of candidates. We find that optimal communication protocols in this setting maximize mutual information between the inputs and messages, as shown in Rita et al. [42].

Supervised discrimination Appendix A.2 explores a variant of the discrimination game that incorporates labels by selecting distractors with labels different from the target. We find that optimal communication strategies in this setting are both diverse (the sender is encouraged to output messages uniformly) and pure (labels distribute with low entropy given a message).

Classification discrimination In Appendix A.3, we consider a format of the discrimination game where the target is excluded from the candidate set, requiring the receiver to identify a candidate that matches the target’s label. We find that optimal solutions in this setup maximize mutual information between messages and labels.

4 Semantic consistency: a prerequisite to meaningful communication

In many EC environments, the message is a latent representation of the input, and is often directly analogous to the corresponding latent vector from continuous training setups (e.g. autoencoding, contrastive learning). That said, the discrete bottleneck in EC offers an additional perspective to the concept of a message; when the message space is smaller than the input space, the communication protocol is forced to be a many-to-one mapping, meaning that each message represents a *set* of inputs.

We consider inputs mapped to the same message as *equivalent* with respect to the communication protocol, thus defining a set of equivalence classes that partitions the input space. With this perspective in mind, we wish to understand whether this partition signals meaningful properties of the inputs. In other words, we ask the question: *Is there a meaningful relationship between inputs mapped to the same message?* This motivation is illustrated in Figure 3.

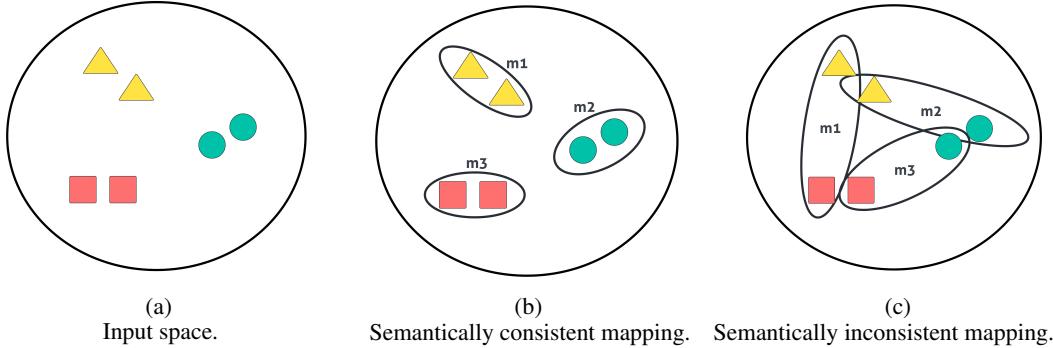


Figure 3: A message describes a set of inputs. Note: the shapes and colors are not part of the input.

The semantics of a message can be interpreted as the set of properties shared across all inputs in its equivalence class. We would like to design a theoretical test that evaluates whether meaningful properties are encoded in this way in an emergent communication protocol. We term this notion *semantic consistency*, based on the idea that inputs with shared properties are more semantically similar, as illustrated in Figure 3. We use the same idea to develop a formal definition: We define a communication protocol to be semantically consistent if a random pair of inputs mapped to the

same message is, in expectation, more similar than two completely random inputs. Formally, for a communication protocol S_θ , consider the following inequality:³

$$\mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \mid S_\theta(x_1) = S_\theta(x_2) \right] < \mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \right]. \quad (1)$$

Equation (1) is simplified in Appendix C.1 into the following formal definition:

Definition 3. A communication protocol S_θ is *semantically consistent* if

$$\mathbb{E}_{m \sim S_\theta(X)} [\text{Var}[X \mid S_\theta(X) = m]] < \text{Var}[X].$$

The difference between the two expressions in Definition 3 is the *explained variance*, given by $\text{Var}_{m \sim S_\theta(X)} [\mathbb{E}[X \mid S_\theta(X) = m]]$. Thus, a communication protocol is semantically consistent if it explains some of the variance in X .

Remark. If the message space is larger than the input space, a sender could map each input to a unique message and lose no information. Any such lossless protocol is trivially semantically consistent.

5 Are EC systems semantically consistent?

Having defined the two common EC environments in Section 3.2, we now apply Definition 3 to assess whether semantically consistent protocols are induced by reconstruction and discrimination tasks. Formally, we answer the following question for each of the tasks:

Is every optimal emergent protocol semantically consistent?

Under the assumption of an unrestricted receiver hypothesis class, we find that:

- In the reconstruction game, the answer is **yes**. We show that equivalent inputs in reconstruction-induced communication protocols are clustered together in input space, matching Figure 3b.
- In the discrimination game, the answer is **no**. The relationship between equivalent inputs in a discrimination-induced protocol can be arbitrary, matching Figure 3c.

All proofs are given in Appendix C.

5.1 Reconstruction results

Ideally, we would like to examine a closed-form set of optimal communication protocols. Unfortunately, there is no closed-form solution to the set of optimal agent pairs in a general reconstruction setting. That said, we do have such a solution to the synchronized receiver, i.e., the optimal receiver for a fixed sender S (the following arg min set contains a single item):

$$R^*(m) = \arg \min_{r \in \mathcal{X}} \mathbb{E}_{x \sim X} [\|r - x\|^2 \mid S(x) = m] = \mathbb{E}[X \mid S(X) = m]$$

By assuming that Φ is unrestricted, we can plug the closed-form synchronized receiver back into the loss function. This yields a transformed objective given in the following lemma.

Lemma 5.1. [proof in page 16] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$\sum_{m \in M} P(S_\theta(X) = m) \cdot \text{Var}[X \mid S_\theta(X) = m] \quad (2)$$

This equivalent objective clearly shows the connection between inputs mapped to the same message: In every optimal solution, these inputs will have high proximity. (In fact, this is the objective function of k-means clustering [36], as elaborated in Appendix B.) Moreover, this formula is the unexplained variance, so minimizing it will maximize the explained variance, leading to the following theorem.

Theorem 5.2. [proof in page 17] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game. Assuming Φ is unrestricted and Θ contains at least one semantically consistent protocol, every optimal communication protocol is semantically consistent.

³We assume that Euclidean distance in the input space entails semantic information. In some types of data, like raw image pixels, using pretrained embeddings may improve the validity of that assumption.

5.2 Discrimination results

In a similar fashion to the reconstruction setting, we have a closed form solution to the synchronized discrimination receiver:⁴

$$R^*(m, x_1, \dots, x_d) = \mathbb{P}(t \mid x_1, \dots, x_d, m)$$

Explicitly (see Lemma 5.3's proof):

$$P(R^*(m, x_1, \dots, x_d) = j) = \frac{1}{|\{x_1, \dots, x_d\} \cap S^{-1}(m)|} \cdot \mathbf{1}\{S(x_j) = m\}$$

Applying this receiver, we get the following discrimination loss:⁵

Lemma 5.3. [proof in page 17] *Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a d -candidates discrimination game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:*

$$\sum_{m \in M} P(S_\theta(X) = m) \cdot \mathbb{E} \log \left(1 + \text{Binomial}(d - 1, P(S_\theta(X) = m)) \right) \quad (3)$$

And when $d = 2$ (a single-distractor game), this simplifies into: $\sum_{m \in M} P(S_\theta(X) = m)^2$.

This equivalent objective reveals the interesting disparity between reconstruction and discrimination. Note that this objective is solely a function of the ‘sizes’ of equivalence classes (the probability for each message). Thus, while the above formula incentivizes the sender to distribute the messages uniformly, it does not impose any constraint on their content; the connection between inputs mapped to the same message could be arbitrary. In formal terms, we prove the following corollary:

Corollary 5.4. [proof in page 19] *If the set of possible messages is finite, any sender agent that assigns them such that their distribution is uniform can be paired with some receiver to obtain a globally optimal loss for the discrimination game.*

Using this corollary, we prove that optimal solutions can be inconsistent in the discrimination setting:

Theorem 5.5. [proof in page 20] *There exists a discrimination game $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ where Φ is unrestricted and Θ contains at least one semantically consistent protocol, in which not all of the optimal communication protocols are semantically consistent.*

6 A missing dimension: spatiality in the message space

While the notion of semantic consistency successfully differentiates between the two EC frameworks, it has a major shortcoming: it does not take into account distances between messages. Spatiality in the message space is integral to concepts like compositionality [5] and ease of teaching [33]. A hypothetical communication protocol that maps similar messages to very different meanings may satisfy Definition 3, but is fundamentally different from natural language. We now present a stricter version of semantic consistency, *spatial meaningfulness*, which does consider distances in the message space, and therefore better indicates an inherent similarity to natural language. In addition, we eliminate the assumption that the receiver’s hypothesis class is unrestricted, in favor of simplicity and non-degeneracy conditions, which are more realistic in the context of natural language.

To introduce distance between messages, we replace the message equality with similarity in Equation (1). Let ε_0 denote a threshold under which messages are considered similar. To avoid reverting back to Definition 3, we require that ε_0 be greater than or equal to ε_M , defined as the minimal distance between messages: $\varepsilon_M := \min_{m_1 \neq m_2 \in M} \|m_1 - m_2\|$.⁶

Definition 4. For $\varepsilon_0 \geq \varepsilon_M$, a communication protocol S_θ is ε_0 -*spatially meaningful* if $\forall 0 < \varepsilon \leq \varepsilon_0$

$$\mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2 \mid \|S_\theta(x_1) - S_\theta(x_2)\| \leq \varepsilon] < \mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2]$$

A communication protocol S_θ is *spatially meaningful* if this definition holds for $\varepsilon_0 = \varepsilon_M$.

⁴Negative log-likelihood is a proper scoring rule; see Lemma 5.3's proof.

⁵The expectation is taken with respect to the binomial distribution.

⁶This is well defined due to our assumption in Section 3.

Note that this definition is stricter than semantic consistency. Informally, Definition 4 requires the protocol to be semantically consistent when combining pairs of close messages.

With this new definition, we are able to reach similar results to the original semantic consistency:

- In the reconstruction game, for any receiver that satisfies two conditions, we show that every synchronized sender is spatially meaningful.
- In the discrimination game, given the same conditions, we show that a synchronized sender is not necessarily semantically consistent (and hence nor spatially meaningful).

We now define two conditions on the receiver.

Simplicity constraint. The first condition limits the receiver’s complexity, and introduces the effect of distance between messages. We do that by bounding the rate of change in receiver’s output, similarly to a Lipschitz constant. The bounding term depends on the variance of the input distribution and the chosen distance threshold. Formally:

Definition 5. Receiver R_φ is (X, M, ε_0) -simple if $\forall x_1, x_2 \in \text{domain}(R_\varphi)$:

$$\|R_\varphi(x_1) - R_\varphi(x_2)\| \leq k \cdot \|x_1 - x_2\|, \quad \text{where } k = \frac{\sqrt{2} - 1}{2\varepsilon_0} \sqrt{\text{Var}[X]}$$

Non-degeneracy constraint. The second condition prevents the receiver from being too weak. After applying the first condition, we no longer have a closed-form solution for the synchronized receiver. Instead, we take a non-optimal receiver, but add a condition to ensure it is strictly better than any constant function. Formally, let φ^C be a receiver that always outputs C , and let $C^* = \arg \min_{x \sim X} \mathbb{E} \ell(\theta, \varphi^C, x)$ be the optimal such constant. (Note the sender does not matter here.)

Definition 6. Receiver R_φ is non-degenerate if

$$\sup_{x \in \mathcal{X}} \ell(\theta^*, \varphi, x) \leq \frac{1}{4} \cdot \mathbb{E}_{x \sim X} \ell(\theta, \varphi^{C^*}, x)$$

for every sender θ^* that is synchronized with R_φ .

6.1 Results

With these definitions, we have the following theorems:

Theorem 6.1. [proof in page 21] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game, let $\varepsilon_0 \geq \varepsilon_M$ and let $\varphi \in \Phi$ such that R_φ is (X, M, ε_0) -simple and non-degenerate. Every sender that is synchronized with R_φ is ε_0 -spatially meaningful.

Theorem 6.2. [proof in page 22] There exists a discrimination game $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$, $\varepsilon_0 \geq \varepsilon_M$ and a receiver $\varphi \in \Phi$ which is (X, M, ε_0) -simple and non-degenerate, where a synchronized sender matching R_φ is not ε -spatially meaningful for any ε .

7 Experiments

To support our theoretical findings, we run experiments on two datasets: (i) the MNIST dataset [30] contains images of a single hand-written digit; (ii) the Shapes dataset [27] contains images of an object with random shape, color and position. We train reconstruction and discrimination (40 distractors) agents using a communication channel of vocabulary size 10 and message length 4 (no EOS token), optimized with Gumbel-Softmax [20, 22]. In the MNIST experiments, we use the digit labels to train another set of agents with the supervised discrimination task (40 distractors). Further data and training details are found in Appendix D.1.

7.1 Semantic consistency, compositionality and task performance

To evaluate semantic consistency empirically, we develop the *empirical message variance* measure (in Appendix D.2) which calculates the empirical variance of each equivalence class and takes a

²Our code is available at <https://github.com/Rotem-BZ/SemanticConsistency>.

weighted average. We measure task performance using discrimination accuracy over 41 candidates (40 distractors), where the referential prediction of reconstruction-trained agents is defined as the closest candidate to the reconstructed target. We also report Topographic Similarity [5], the most common compositionality measure from literature. See Appendix E.1 for results with other compositionality measures (PosDis [8], BosDis [8] and Speaker-PosDis [13]). Additional information on each metric is given in Appendix D.2.

Table 1 and Table 2 show results over the Shapes and MNIST validation sets, respectively (showing standard deviation). While all setups achieve good task performance over MNIST, the discrimination game yields higher discrimination accuracy over the more difficult Shapes dataset. As expected. Interestingly, the discrimination-trained agents utilize less of the communication channel’s capacity, as shown by the number of unique messages in both tables. This presents a gap from the theoretical analysis, as the discrimination setting should benefit from maximal message diversity (this can be seen in Lemma 5.3). To enable a fair comparison, we establish individual random baselines that maintain the number of inputs mapped to each message but randomize their assignment. Thus, the baseline’s message variance shows the level of message diversity induced by the task, whereas the improvement over the baseline indicates semantic consistency. Over both datasets, the reconstruction task induces significantly more semantically consistent protocols both in absolute value and improvement over the baseline, as predicted by our theoretical findings. Furthermore, TopSim strongly favors the reconstruction setting, suggesting a connection between semantic consistency and compositionality. We note that the difference measured by TopSim is more significant over the Shapes dataset, which is itself more compositional, as objects have several attributes. These results, and the correlations reported in Appendix E.1, indicate little to no correlation between the evaluated compositionality and the discrimination objective performance. In fact, within the set of reconstruction runs on Shapes, their correlation is negative (-0.24). This illustrates the counter-intuitive nature of this objective, which we discuss in Section 5.2.

Table 1: Empirical results on Shapes, averaged over five randomly initialized training runs.

EC setup	Unique Msgs	Disc. accuracy \uparrow	TopSim \uparrow	Msg Var \downarrow	
				Trained	Rand
Reconstruction	306.60 ± 28.52	31.64 ± 2.51	0.34 ± 0.02	1334.38 ± 78.05	2554.77 ± 108.19
Discrimination	251.60 ± 29.53	61.96 ± 4.78	0.09 ± 0.01	2280.24 ± 157.38	2793.65 ± 115.45

Table 2: Empirical results on MNIST, averaged over three randomly initialized training runs.

EC setup	Unique Msgs	Disc. accuracy \uparrow	TopSim \uparrow	Msg Var \downarrow	
				Trained	Rand
Reconstruction	2523.66 ± 30.0	88.00 ± 0.6	0.365 ± 0.042	371.63 ± 2.1	1029.57 ± 13.9
Discrimination	402.00 ± 70.4	78.76 ± 10.4	0.360 ± 0.036	1226.14 ± 42.4	1784.59 ± 24.3
Supervised disc.	287.33 ± 28.5	87.10 ± 5.4	0.269 ± 0.044	1381.72 ± 41.6	1821.53 ± 12.2

7.2 Message purity

Since the MNIST dataset has only the digit label, we expect meaningful communication protocols to signal digits in their messages. To evaluate this, we calculate the percentage of images that have the majority digit label of their equivalence class. This can be interpreted as the highest attainable accuracy of a classifier mapping messages to digits. Figure 4 shows the most significant improvement over the baseline in the supervised discrimination game, as expected by Lemma A.2. The reconstruction task induces a nearly perfect digit description protocol as well, despite being unsupervised, providing evidence of its ability to unveil meaningful properties, in agreement with our findings. On the other hand, message purity is much lower in the unsupervised discrimination setting, again showcasing that task success does not guarantee an intuitive communication strategy.

In Appendix E.2, we perform a similar analysis over the Shapes dataset, by evaluating message purity with respect to each attribute individually or to a maximum aggregation. We find similar but weaker results.

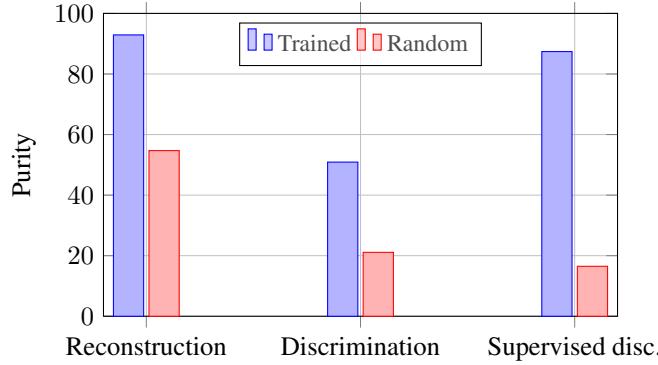


Figure 4: Average message purity, comparing trained models to random baselines.

7.3 Spatial meaningfulness analysis

We develop an evaluation method for spatial meaningfulness, which we term *cluster variance*, based on the message variance measure. While we do not find a decisive result using this method, we suggest several ideas for further investigation; see Appendix E.3 for the full analysis.

8 Limitations

Our analysis compares the two frameworks via their optimal solutions, and often assumes unrestricted hypothesis classes. In reality, the complexity of the agents is limited by the chosen architectures. Optimization-error affects the emergent protocol as well, especially when it is discrete, as regular differentiable training is not possible. Future work may further analyze these assumptions from the theoretical side, or evaluate their effect empirically. This limitation is evident in our empirical results (Section 7), where channel utilization in the discrimination game behaves differently than expected.

Our main results regarding the discrimination game deal with a simplistic version, where candidates are chosen independently. While we analyze some common variations in the appendix, many other ideas have been proposed and shown to have significant benefits, including multiple-target games [6, 37], different-view candidate representations [10, 16], and multi-modality [14].

Our results assume, as illustrated in Figure 3, that proximity in the input space entails semantic information, and specifically that Euclidean distance broadly indicates the level of semantic similarity. Future work may investigate other distance metrics or evaluate the validity of this assumption.

9 Conclusion

Based on properties of natural language, we have defined notions of semantic consistency and spatial meaningfulness, meant to evaluate meaningfulness in emergent communication protocols. Using these definitions, we found that communication protocols generated to solve the reconstruction objective have messages with inherent meaning, while the discrimination objective can be solved by unintuitive systems. Our findings provide insight into known empirical results in EC, such as the ability of agents in the discrimination game to perform well on unseen random noise. The theoretical tools that we have proposed can be used for future investigation and design of EC environments. As a main takeaway, we conclude that distance-based EC environments have promising potential in the prospect of inducing characteristics of natural language, whereas probability-based objectives must be more carefully designed.

Acknowledgements

This research was supported by grant no. 2022330 from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel, the Israel Science Foundation (grant no. 448/20), an Azrieli Foundation Early Career Faculty Fellowship, and an AI Alignment grant from Open Philanthropy. OP was funded by Project CETI via grants from Dalio Philanthropies and Ocean X; Sea Grape Foundation; Virgin Unite and Rosamund Zander/Hansjorg Wyss through The Audacious Project: a collaborative funding initiative housed at TED.

References

- [1] Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2018.
- [2] Michal Auersperger and Pavel Pecina. Defending compositionality in emergent languages. *NAACL 2022*, page 285, 2022.
- [3] Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, 2018.
- [4] Nicolo' Brandizzi. Toward more human-like ai communication: A review of emergent communication research. *IEEE Access*, 11:142317–142340, 2023.
- [5] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.
- [6] Boaz Carmeli, Yonatan Belinkov, and Ron Meir. Concept-best-matching: Evaluating compositionality in emergent communication. *arXiv preprint arXiv:2403.14705*, 2024.
- [7] Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*, 32, 2019.
- [8] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, 2020.
- [9] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.
- [10] Edward Choi, Angeliki Lazaridou, and Nando De Freitas. Compositional obverter communication learning from raw visual input. In *6th International Conference on Learning Representations, ICLR 2018*. International Conference on Learning Representations, ICLR, 2018.
- [11] Kevin Denamganaï and James Alfred Walker. On (emergent) systematic generalisation and compositionality in visual referential games with straight-through gumbel-softmax estimator. *arXiv preprint arXiv:2012.10776*, 2020.
- [12] Kevin Denamganaï and James Alfred Walker. Referentialgym: A nomenclature and framework for language emergence & grounding in (visual) referential games. *arXiv preprint arXiv:2012.09486*, 2020.
- [13] Kevin Denamganaï, Sondess Missaoui, and James Alfred Walker. Visual referential games further the emergence of disentangled representations. *arXiv preprint arXiv:2304.14511*, 2023.
- [14] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. Emergent communication in a multi-modal, multi-step referential game. In *International Conference on Learning Representations*, 2018.
- [15] Shangmin Guo, Yi Ren, Kory Wallace Mathewson, Simon Kirby, Stefano V Albrecht, and Kenny Smith. Expressivity of emergent languages is a trade-off between contextual complexity and unpredictability. In *International Conference on Learning Representations*, 2021.
- [16] Kamal Gupta, Gowthami Somepalli, Anubhav Anubhav, Vinoj Yasanga Jayasundara Magalle Hewa, Matthias Zwicker, and Abhinav Shrivastava. Patchgame: Learning to signal mid-level patches in referential games. *Advances in Neural Information Processing Systems*, 34:26015–26027, 2021.
- [17] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30, 2017.

- [18] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
- [19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.
- [20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview.net, 2017.
- [21] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.
- [22] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Egg: a toolkit for research on emergence of language in games. *EMNLP-IJCNLP 2019*, page 55, 2019.
- [23] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Entropy minimization in emergent languages. In *International Conference on Machine Learning*, pages 5220–5230. PMLR, 2020.
- [24] Tomasz Korbak, Julian Zubek, and Joanna Rączaszek-Leonardi. Measuring non-trivial compositionality in emergent communication. *arXiv preprint arXiv:2010.15058*, 2020.
- [25] Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge ‘naturally’ in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, 2017.
- [26] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2013.
- [27] Alexander Kuhnle and Ann Copestake. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*, 2017.
- [28] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*, 2016.
- [29] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*, 2018.
- [30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [31] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Counteracting language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4385–4395, 2019.
- [32] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.
- [33] Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. *Advances in neural information processing systems*, 32, 2019.
- [34] Toru Lin, Jacob Huh, Christopher Stauffer, Ser Nam Lim, and Phillip Isola. Learning to ground multi-agent communication with autoencoders. *Advances in Neural Information Processing Systems*, 34:15230–15242, 2021.
- [35] Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 693–701, 2019.
- [36] JB MACQUEEN. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics & Probability*, volume 1, pages 281–297. University of California Press, 1967.
- [37] Jesse Mu and Noah Goodman. Emergent communication of generalizations. *Advances in Neural Information Processing Systems*, 34:17994–18007, 2021.
- [38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

- [39] Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Improving image autoencoder embeddings with perceptual loss. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.
- [40] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *8th International Conference on Learning Representations*, 2020.
- [41] Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. “lazimpa”: Lazy and impatient neural agents learn to communicate efficiently. In *CONLL 2020-The SIGNLL Conference on Computational Natural Language Learning*, 2020.
- [42] Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. Emergent communication: Generalization and overfitting in lewis games. *Advances in Neural Information Processing Systems*, 35:1389–1404, 2022.
- [43] Shane Steinert-Threlkeld, Xuhui Zhou, Zeyu Liu, and CM Downey. Emergent communication fine-tuning (ec-ft) for pretrained language models. In *Emergent Communication Workshop at ICLR 2022*, 2022.
- [44] Mycal Tucker, Roger Levy, Julie A Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. *Advances in neural information processing systems*, 35: 22214–22228, 2022.
- [45] Mycal Tucker, Roger P Levy, Julie Shah, and Noga Zaslavsky. Generalization and translatability in emergent communication via informational constraints. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022.
- [46] Shunyu Yao, Mo Yu, Yang Zhang, Karthik Narasimhan, Joshua Tenenbaum, and Chuang Gan. Linking emergent and natural languages via corpus transfer. In *International Conference on Learning Representations*, 2022.

A Analysis of discrimination variations

In this section, we extend our analysis of the discrimination game to three common variations from EC literature.

A.1 Global discrimination

In this variation, rather than sampling a small set of candidates, the receiver must discriminate between the entire set of data points, which does not have to be finite nor countable. In related work this game is often referred to as the reconstruction game [7, 41, 42, 44], as it can be interpreted as a generative task. However, since it uses the negative log-likelihood objective and not a distance-based one, we call it *global discrimination*. Note that when the set of data inputs \mathcal{X} is finite, this is similar to a special case of the regular discrimination game, where the number of candidates d is set to $|\mathcal{X}|$ (and candidates are sampled without replacement). This equivalence is also mentioned by [12]. Denote:

- $\Phi_{\text{global}}^{M, \mathcal{X}}$ to be the set of all functions of the form $R: M \rightarrow \Delta(\mathcal{X})$.
- $\ell_{\text{global}}(\theta, \varphi, x) := -\log P(R_\varphi(S_\theta(x)) = x)$

Definition 7. An EC setup $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ is a *global discrimination game* if $\Phi \subseteq \Phi_{\text{global}}^{M, \mathcal{X}}$ and $\ell = \ell_{\text{global}}$.

Lemma A.1. [proof in page 24] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a global discrimination game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$-I(X; S_\theta(X))$$

We see that optimal protocols maximize mutual information between inputs and messages. This corresponds to a popular idea in representation learning literature [19, 38]. However, similarly to the objective shown in Lemma 5.3 for the regular discrimination game, this communication goal does not consider distances between inputs. In fact, mutual information can be calculated for non-numerical input spaces.

A.2 Supervised discrimination

In this variation, labels are incorporated into the game via the candidate choice mechanism. The distractors (candidates that are not the target) are guaranteed to not have the same label as the target. This means that the receiver will never have to tell apart two inputs from the same class.

Let \mathcal{Y} be a finite set of labels, and let $\text{label} : \mathcal{X} \rightarrow \mathcal{Y}$ be a (deterministic) function mapping input to label. Denote the distribution over labels $Y = \text{label}(X)$. We assume that this random variable has a uniform distribution, i.e., that the classes are balanced. The supervised variation uses the same receiver functions as in Definition 2, and a slightly different loss function:

$$\ell_{\text{supervised}}^{X, Y, d}(\theta, \varphi, x) := \mathbb{E}_{\substack{t \sim U\{1, \dots, d\}, x_t=x \\ x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d \sim P(X|Y \neq \text{label}(x))^{d-1}}} -\log P(R_\varphi(S_\theta(x), x_1, \dots, x_d) = t)$$

Definition 8. For $d \leq |\mathcal{Y}|$, an EC setup $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ is a d -candidates supervised discrimination game if $\Phi \subseteq \Phi_{\text{discrimination}}^{M, \mathcal{X}, d}$ and $\ell = \ell_{\text{supervised}}^{X, Y, d}$.

Lemma A.2. [proof in page 24] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a 2-candidates supervised discrimination game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$\sum_{m \in M} P(S_\theta(X) = m)^2 - \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(S_\theta(X) = m, \text{label}(X) = y)^2$$

This result is interpretable too. The first expression is the unsupervised objective that encourages uniformity of message probabilities, which we encountered in Lemma 5.3. The second expression encourages non-diversity of labels within each equivalence class (optimal if all inputs mapped to a message share the same label). This means that optimal messages would have high mutual information with the labels, which in turn could introduce semantic consistency. However, within each class, the communication protocol would not necessarily be able to distinguish between inputs; perfect accuracy can be achieved with just $|\mathcal{Y}|$ unique messages.

This result presents another key implication, regarding complexity in the discrimination game [15]. We can consider a candidate choice mechanism that makes sure all distractors have the same label as the target (the opposite of the classic supervised game). This task will require the receiver to tell apart similar inputs, resulting in high contextual complexity. Interestingly, this task incentivizes the sender to map same-label inputs to different messages, in sharp contrast from Lemma A.2. Therefore, messages learned to solve this high-complexity task will have **low mutual information** with the labels, i.e., this game is likely to result in counter-intuitive communication protocols.

A.3 Classification discrimination

The classification variation builds on top of the supervised version, by swapping out the target from the set of candidates, and replacing it with a random input that has the same label. In other words, the receiver attempts to guess the candidate with the same label as the target. This game is structured such that every class has exactly one representation in the set of candidates. This is somewhat related to the object-focused referential game [10, 12].

Let $\mathcal{Y} = \{1, \dots, n\}$ be a finite set of labels, and let $\text{label} : \mathcal{X} \rightarrow \mathcal{Y}$ be a (deterministic) function mapping input to label. Denote the distribution over labels $Y = \text{label}(X)$. The classification variation uses the same receiver functions as in Definition 2, and the following loss function:

$$\ell_{\text{classification}}^{X,Y}(\theta, \varphi, x) := \mathbb{E}_{\{x_i \sim P(X|Y=i)\}_{i=1}^n} - \log P(R_\varphi(S_\theta(x), x_1, \dots, x_n) = \text{label}(x))$$

Definition 9. An EC setup $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ is a *classification discrimination game* if $\Phi \subseteq \Phi_{\text{discrimination}}^{M, \mathcal{X}, n}$ and $\ell = \ell_{\text{classification}}^{X,Y}$.

Note that if \mathcal{X} is finite, we could choose $n = |\mathcal{X}|$ and $\text{label}(x_i) = i$, i.e. define labels as the identity mapping, and end up with the global discrimination game.

Lemma A.3. [proof in page 25] Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a classification discrimination game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$-I(Y; S_\theta(X))$$

Optimal communication protocols in the classification game maximize mutual information with the labels. If the labels signal meaningful properties of the inputs, protocol induced by this game will reflect that, and be semantically consistent. However, similarly to the supervised discrimination game, optimal solutions can use just $|\mathcal{Y}|$ different messages, as they do not have to describe any properties of the input other than its label.

B Reconstruction and k-means clustering

Lemma 5.1 shows that optimal communication protocols in the reconstruction setting minimize the objective function of k-means clustering [36], where each equivalence class corresponds to a cluster. In this section we show a stronger connection between the two environments. Namely, we show that if X is a uniform distribution with finite support, both Θ and Φ are unrestricted, and the message space M is finite, the reconstruction game can simulate k-means clustering with $k = |M|$ (the equivalence classes correspond to clusters). This happens if we alternate between each agent's optimization:

Assignment step When the receiver is frozen and sender is optimized, the best solution (synchronized sender) assigns each input to its projection on Receiver's Image:

$$R(S^*(x)) = R(\arg \min_{m \in M} \|x - R(m)\|) = \arg \min_{x' \in \text{Img}(R)} \|x - x'\|.$$

In other words, the sender chooses the message (cluster) with the closest receiver output (centroid) to its input.

Update step When the sender is frozen and receiver is optimized, the best solution (synchronized receiver) sets each output to be the mean over the relevant inputs:

$$R^*(m) = \arg \min_{r \in \mathcal{X}} \mathbb{E}_{x \sim I} [\|r - x\|^2 \mid S(x) = m] = \mathbb{E}[X \mid S(X) = m]$$

In other words, the receiver updates its output (centroid) to be the mean over inputs mapped to the given message (cluster).

C Proofs

For ease of notation, we often write S, S^*, R, R^* instead of $S_\theta, S_{\theta^*}, R_\varphi, R_{\varphi^*}$ when the hypotheses are clear from context. Additionally, we use square brackets as a simple notation of a message's equivalence class, i.e.,

$$[m] \equiv [m]_\theta := S_\theta^{-1}(m) = \{x \in \mathcal{X} : S_\theta(x) = m\}$$

C.1 Simplification of the semantic consistency definition

simplification of Equation (1). Recall Equation (1):

$$\mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \mid S_\theta(x_1) = S_\theta(x_2) \right] < \mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \right]$$

Note that since \mathcal{X} is bounded, these values as well as the expectation and variance of X are finite. Using the identity $\mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2] = 2\text{Var}[X]$, the left term is equal to

$$\begin{aligned} & \mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \mid S_\theta(x_1) = S_\theta(x_2) \right] \\ &= \mathbb{E}_{m \sim S_\theta(X)} \mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \mid S_\theta(x_1) = S_\theta(x_2), S_\theta(x_1) = m \right] \\ &= \mathbb{E}_{m \sim S_\theta(X)} \mathbb{E}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \mid S_\theta(x_1) = m, S_\theta(x_2) = m \right] \\ &= 2 \cdot \mathbb{E}_{m \sim S_\theta(\mathcal{X})} [\text{Var}[X \mid S_\theta(X) = m]] \end{aligned}$$

and the right term is simply $2\text{Var}[X]$. Applying this to Equation (1) yields Definition 3. \square

C.2 Semantic consistency proofs

Reconstruction

Lemma 5.1. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$\sum_{m \in M} P(S_\theta(X) = m) \cdot \text{Var}[X \mid S_\theta(X) = m]$$

Proof of Lemma 5.1. Let $X_{[m]}$ denote the distribution $P(X \mid X \in [m])$. The synchronized receiver for sender S is defined by:

$$R^* = \arg \min_{R \in \Phi} \mathbb{E}_{x \sim X} \|R(S(x)) - x\|^2$$

Since Φ is assumed to be unrestricted, we get

$$\begin{aligned} R^*(m) &= \arg \min_{r \in \mathcal{X}} \mathbb{E}_{x \sim X} [\|r - x\|^2 \mid S(x) = m] \\ &= \arg \min_{r \in \mathcal{X}} \mathbb{E}_{x \sim X_{[m]}} \|r - x\|^2 \\ &= \mathbb{E}[X_{[m]}] \end{aligned}$$

We can apply this synchronized receiver back in the reconstruction loss function without changing the set of optimal sender hypotheses, since a sender is optimal if and only if when paired with a

synchronized receiver, the pair is optimal. Note that $[m] = S^{-1}(m)$ depends on sender. We get:

$$\begin{aligned}
S^* &\in \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \|R^*(S(x)) - x\|^2 \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \|\mathbb{E}[X_{[S(x)]}] - x\|^2 \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \left[P(X \in [m]) \cdot \mathbb{E}_{x \sim X^{[m]}} \|\mathbb{E}[X_{[S(i)]}] - x\|^2 \right] \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \left[P(X \in [m]) \cdot \mathbb{E}_{x \sim X^{[m]}} \|\mathbb{E}[X_{[m]}] - x\|^2 \right] \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \cdot \text{Var}[X_{[m]}]
\end{aligned}$$

The third line is the law of total expectation, where the events are the assignment to message m . \square

Theorem 5.2. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game. Assuming Φ is unrestricted and Θ contains at least one semantically consistent protocol, every optimal communication protocol is semantically consistent.

Proof of Theorem 5.2. Following Lemma 5.1, let θ^* be an optimal reconstruction sender, i.e.,

$$\begin{aligned}
\theta^* &\in \arg \min_{\theta} \sum_{m \in M} P(X \in [m]_{\theta}) \cdot \text{Var}[X \mid X \in [m]_{\theta}] \\
&= \arg \min_{\theta} \mathbb{E}_{\theta} [\text{Var}[X \mid S_{\theta}(X)]] \\
&= \arg \max_{\theta} -\mathbb{E}_{\theta} [\text{Var}[X \mid S_{\theta}(X)]] \\
&= \arg \max_{\theta} \text{Var}[X] - \mathbb{E}_{\theta} [\text{Var}[X \mid S_{\theta}(X)]]
\end{aligned}$$

The final formula is non-negative by variance decomposition. Since Θ contains a semantically consistent protocol, this expression can be greater than zero, so since θ^* maximizes it, we get

$$\text{Var}[X] - \mathbb{E}_{\theta} [\text{Var}[X \mid S_{\theta}(X)]] > 0$$

Which proves that Definition 3 is satisfied. \square

Discrimination

Lemma 5.3. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a d -candidates discrimination game. Assuming Φ is unrestricted, a sender S_{θ} is optimal if and only if it minimizes the following objective:

$$\sum_{m \in M} P(S_{\theta}(X) = m) \cdot \mathbb{E} \log \left(1 + \text{Binomial}(d-1, P(S_{\theta}(X) = m)) \right)$$

And when $d = 2$ (a single-distractor game), this simplifies into

$$\sum_{m \in M} P(S_{\theta}(X) = m)^2$$

Proof of Lemma 5.3. Given message m and candidates (x_1, \dots, x_d) , such that t is the index of the target ($S(x_t) = m$), denote $C = \{x_1, \dots, x_d\} \cap [m]$. Since t is sampled uniformly and only candidates in C have a non-zero likelihood of being the target, we have for all $j \in \{1, \dots, d\}$

$$\begin{aligned}
P(t = j \mid x_1, \dots, x_d, m) &= \frac{f_X(x_1, \dots, x_d) \cdot P(t = j \mid x_1, \dots, x_d) \cdot P(S(x_t) = m \mid x_1, \dots, x_d, t = j)}{f_X(x_1, \dots, x_d) \cdot P(S(x_t) = m \mid x_1, \dots, x_d)} \\
&= \frac{\frac{1}{d} \cdot \mathbf{1}\{S(x_j) = m\}}{\frac{|C|}{d}} \\
&= \frac{1}{|C|} \cdot \mathbf{1}\{x_j \in C\}
\end{aligned}$$

Denote this probability p_j . A synchronized receiver for sender S is defined as

$$\begin{aligned} R^* &\in \arg \min_{R \in \Phi} \mathbb{E}_{x \sim X} \ell(\theta, R, x) \\ &= \arg \min_{R \in \Phi} \mathbb{E}_{x \sim X} \mathbb{E}_{\substack{t \sim U\{1, \dots, d\}, x_t=x \\ x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d \sim X^{d-1}}} -\log P(R(S(x), x_1, \dots, x_d) = t) \\ &= \arg \min_{R \in \Phi} \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} -\log P(R(S(x_t), x_1, \dots, x_d) = t) \end{aligned}$$

Since Φ is assumed to be unrestricted, we get

$$\begin{aligned} R^*(m, x_1, \dots, x_d) &\in \arg \min_{(q_1, \dots, q_d) \in \Delta\{1, \dots, d\}} \mathbb{E}_{t \sim U\{1, \dots, d\}} [-\log q_t \mid x_1, \dots, x_d, m, S(x_t) = m] \\ &= \arg \min_{(q_1, \dots, q_d) \in \Delta\{1, \dots, d\}} -\sum_{j=1}^d p_j \log q_j \end{aligned}$$

A lower bound on the objective (Gibbs' inequality):

$$-\sum_{j=1}^d p_j \log q_j \geq -\sum_{j=1}^d p_j \log p_j$$

This lower bound is achieved by selecting $q_j = p_j$ (negative log-likelihood is a proper scoring rule). Note that $p_j \geq 0 \forall j$ and $\sum_{j=1}^d p_j = 1$, making it a legal output for R . We now apply this receiver back in the loss function, similarly to the proof of Lemma 5.1. Recall that x_1, \dots, x_d are the candidates, and given the target's index t , the other candidates $(x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d)$ are called distractors. Also note that $[m] = S^{-1}(m)$ depends on sender.

$$\begin{aligned} S^* &\in \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \ell(S, R^*, x) \\ &= \arg \min_{S \in \Theta} \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} -\log P(R^*(S(x_t), x_1, \dots, x_d) = t) \\ &= \arg \min_{S \in \Theta} \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} -\log \frac{1}{|\{x_1, \dots, x_d\} \cap [S(x_t)]|} \\ &= \arg \min_{S \in \Theta} \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} \log |\{x_1, \dots, x_d\} \cap [S(x_t)]| \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} [\log |\{x_1, \dots, x_d\} \cap [S(x_t)]| \mid S(x_t) = m] \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \mathbb{E}_{\substack{x_1, \dots, x_d \sim X \\ t \sim U\{1, \dots, d\}}} \log \left(1 + |\{x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d\} \cap [m]| \right) \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \mathbb{E}_{\text{distractors} \sim X^{(d-1)}} \log \left(1 + |\text{distractors} \cap [m]| \right) \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \cdot \mathbb{E} \log \left(1 + \text{Binomial}(d-1, P(X \in [m])) \right) \end{aligned}$$

if $d = 2$ (a single-distractor game), the expression simplifies to:

$$\begin{aligned} &\arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \cdot \mathbb{E} \log \left(1 + \text{Binomial}(1, P(I \in [m])) \right) \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \cdot \left(P(X \in [m]) \cdot \log 2 + (1 - P(X \in [m])) \cdot \log 1 \right) \\ &= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m])^2 \end{aligned}$$

□

Corollary 5.4. If the set of possible messages is finite, any sender agent that assigns them such that their distribution is uniform can be paired with some receiver to generate globally optimal loss for the discrimination game.

Proof of Corollary 5.4. Let $n = |M|$ be the finite number of available messages. Denote the vector $\mathbf{p} = (p_1, \dots, p_n)$ such that $p_i = P(X \in [m_i])$. For $x \in [0, 1]$, denote the function

$$f(x) = x \cdot \mathbb{E} \log(1 + \text{Binomial}(d - 1, x))$$

Following Lemma 5.3, we can formulate the task as a constrained optimization problem on the simplex:

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}^n} \quad & F(\mathbf{p}) := \sum_{i=1}^n f(p_i) \\ \text{s.t.} \quad & \mathbf{p}^T \mathbf{1} = 1 \\ & \mathbf{p} \geq 0 \end{aligned}$$

Note that the distribution of X may create additional constraints. We will prove that f is convex, meaning that this is a convex problem, so any KKT solution is global minimum [26]. The uniform solution $\mathbf{p}^* = (\frac{1}{n}, \dots, \frac{1}{n})$ is indeed KKT: assigning $\lambda = \vec{0}$ and $\mu = -\nabla f(\frac{1}{n})$ we have

$$\nabla F(\mathbf{p}^*) + \lambda^T \nabla(-\mathbf{p})(\mathbf{p}^*) + \mu \cdot \nabla(\mathbf{p}^T \mathbf{1} - 1)(\mathbf{p}^*) = \nabla f\left(\frac{1}{n}\right) + \lambda^T(-I_n) + \mu \cdot \mathbf{1} = \vec{0}$$

It is left to prove that f is convex over $[0, 1]$. For simplicity, we replace $d - 1$ with $d \in \mathbb{N}^+$:

$$f(x) = x \cdot \mathbb{E}[\log(1 + \text{Binom}(d, x))]$$

This convexity proof is adjusted from mathoverflow ([link](#)). Denote

$$\tilde{g}(k) = \log(1 + k)$$

and

$$B_g^d(x) = \mathbb{E}[g(\text{Binom}(d, x))] = \sum_{k=0}^d g(k) \binom{d}{k} x^k (1-x)^{d-k} \Rightarrow f(x) = x \cdot B_g^d(x)$$

We have

$$\begin{aligned} B_g^{d'}(x) &= \sum_{k=0}^d g(k) \binom{d}{k} k x^{k-1} (1-x)^{d-k} - \sum_{k=0}^d g(k) \binom{d}{k} (d-k) x^k (1-x)^{d-k-1} \\ &= \sum_{k=1}^d g(k) \binom{d}{k} k x^{k-1} (1-x)^{d-k} - \sum_{k=0}^{d-1} g(k) \binom{d}{k} (d-k) x^k (1-x)^{d-k-1} \end{aligned}$$

Using the identities $\binom{d}{k} k = d \binom{d-1}{k-1}$ and $\binom{d}{k} (d-k) = d \binom{d-1}{k}$, we get

$$\begin{aligned} B_g^{d'}(x) &= \sum_{k=1}^d g(k) d \binom{d-1}{k-1} x^{k-1} (1-x)^{d-k} - \sum_{k=0}^{d-1} g(k) d \binom{d-1}{k} x^k (1-x)^{d-k-1} \\ &= d \sum_{k=0}^{d-1} g(k+1) \binom{d-1}{k} x^k (1-x)^{d-1-k} - d \sum_{k=0}^{d-1} g(k) \binom{d-1}{k} x^k (1-x)^{d-1-k} \\ &= d \sum_{k=0}^{d-1} (g(k+1) - g(k)) \binom{d-1}{k} x^k (1-x)^{d-1-k} \\ &= d \cdot B_{\Delta g}^{d-1}(x) \end{aligned}$$

where $\Delta g(k) = g(k+1) - g(k)$.

We can apply this again to get the second derivative of B_g^d :

$$B_g^{d''}(x) = d \cdot B_{\Delta g}^{d-1'}(x) = d(d-1) B_{\Delta^2 g}^{d-2}(x)$$

We can now write the second derivative of f :

$$\begin{aligned} f''(x) &= 2B_{\tilde{g}}^{d'}(x) + x \cdot B_{\tilde{g}}^{d''}(x) \\ &= 2d \cdot B_{\Delta \tilde{g}}^{d-1}(x) + x \cdot d(d-1)B_{\Delta^2 \tilde{g}}^{d-2}(x) \\ &= 2d \sum_{k=0}^{d-1} \Delta \tilde{g}(k) \binom{d-1}{k} x^k (1-x)^{d-1-k} + d(d-1) \sum_{k=0}^{d-2} \Delta^2 \tilde{g}(k) \binom{d-2}{k} x^{k+1} (1-x)^{d-2-k} \end{aligned}$$

Since $\binom{d-2}{k-1} = \frac{k}{d-1} \binom{d-1}{k}$, the second term can be written as

$$\begin{aligned} d(d-1) \sum_{k=0}^{d-2} \Delta^2 \tilde{g}(k) \binom{d-2}{k} x^{k+1} (1-x)^{d-2-k} &= d(d-1) \sum_{k=1}^{d-1} \Delta^2 \tilde{g}(k-1) \binom{d-2}{k-1} x^k (1-x)^{d-1-k} \\ &= d \sum_{k=1}^{d-1} \Delta^2 \tilde{g}(k-1) k \binom{d-1}{k} x^k (1-x)^{d-1-k} \\ &= d \sum_{k=0}^{d-1} \Delta^2 \tilde{g}(k-1) k \binom{d-1}{k} x^k (1-x)^{d-1-k} \end{aligned}$$

Thus, we have

$$\begin{aligned} f''(x) &= 2d \sum_{k=0}^{d-1} \Delta \tilde{g}(k) \binom{d-1}{k} x^k (1-x)^{d-1-k} + d \sum_{k=0}^{d-1} \Delta^2 \tilde{g}(k-1) k \binom{d-1}{k} x^k (1-x)^{d-1-k} \\ &= d \sum_{k=0}^{d-1} (2\Delta \tilde{g}(k) + \Delta^2 \tilde{g}(k-1) k) \binom{d-1}{k} x^k (1-x)^{d-1-k} \end{aligned}$$

To finish the proof, we show that the term $h(k) \equiv 2\Delta \tilde{g}(k) + \Delta^2 \tilde{g}(k-1) k$ is positive, proving that f is convex.

$$\begin{aligned} \Delta^2 g(k) &= \Delta g(k+1) - \Delta g(k) = g(k+2) - g(k+1) - (g(k+1) - g(k)) \\ &= g(k+2) + g(k) - 2g(k+1) \\ 2\Delta g(k) + \Delta^2 g(k-1) k &= 2g(k+1) - 2g(k) + k(g(k+1) + g(k-1) - 2g(k)) \\ &= (k+2) \cdot g(k+1) + k \cdot g(k-1) - 2(k+1) \cdot g(k) \end{aligned}$$

We get

$$h(k) \equiv 2\Delta \tilde{g}(k) + \Delta^2 \tilde{g}(k-1) k = (k+2) \cdot \log(k+2) + k \cdot \log(k) - 2(k+1) \cdot \log(k+1)$$

For $k > 0$, the function $t \log t$ is convex, thus $h(k) > 0$ by Jensen's inequality. Also, $h(0) = 2 \log 2 > 0$.

□

Theorem 5.5. There exists a discrimination game $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ where Φ is unrestricted and Θ contains at least one semantically consistent protocol, in which not all of the optimal communication protocols are semantically consistent.

Proof of Theorem 5.5. If X is a uniform distribution over a finite support, then as shown by Corollary 5.4, a Sender is optimal if it maps the same number of inputs to every message. Thus, for example, if $|\mathcal{X}| = 2 \cdot |M|$ (both finite), any Sender that maps exactly two inputs to each message is optimal, and we can define such an agent by mapping the least similar pairs of inputs together, creating an optimal Sender that does not induce a meaning-consistent latent space. For a concrete example, see the proof for Theorem 6.2. □

C.3 Spatial meaningfulness proofs

Theorem 6.1. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game, let $\varepsilon_0 \geq \varepsilon_M$ and let $\varphi \in \Phi$ such that R_φ is (X, M, ε_0) -simple and non-degenerate. Every sender that is synchronized with R_φ is ε_0 -spatially meaningful.

Proof of Theorem 6.1. In the reconstruction task, the optimal constant receiver is

$$\begin{aligned} C^* &= \arg \min_{x \sim X} \mathbb{E}_{x \sim X} \ell(\theta, \varphi^C, x) \\ &= \arg \min_{x \sim X} \mathbb{E}_{x \sim X} \|C - x\|^2 \\ &= \mathbb{E}_{x \sim X}[x] \equiv \bar{X} \end{aligned}$$

Let θ^* be a sender synchronized with R_φ . Denote $L = \sup_{x \in \mathcal{X}} \ell(\theta^*, \varphi, x)$ and $\bar{L} = \frac{1}{2} \cdot \mathbb{E}_{x \sim X} \|x - \bar{X}\|^2$. Since R_φ is strictly better than constant, we have

$$\forall x \in \mathcal{X} : \ell(\theta^*, \varphi, x) \leq L \leq \frac{\bar{L}}{2}$$

Denote the receiver's prediction $R_\varphi(S_{\theta^*}(x)) := \hat{x}_\varphi$. Note that when the sender's hypothesis class Θ is unrestricted, this becomes the projection on R_φ 's image:

$$R_\varphi(S_{\theta^*}(x)) = R_\varphi(\arg \min_{m \in M} \|x - R_\varphi(m)\|) = \arg \min_{x' \in \text{Img}(R_\varphi)} \|x - x'\| = \text{Proj}_{\text{Img}(R_\varphi)}(x).$$

By the definition of $\ell_{\text{reconstruction}}$, we get

$$\begin{aligned} \forall x \in \mathcal{X} \quad \|x - \hat{x}_\varphi\|^2 &\leq L \\ \Rightarrow \forall x \in \mathcal{X} \quad \|x - \hat{x}_\varphi\| &\leq \sqrt{L} \end{aligned}$$

Let $l_R = \frac{\sqrt{2}-1}{2\varepsilon_0} \sqrt{\text{Var}[X]}$ be the constant from Definition 5. For any $\varepsilon > 0$ and $x_1, x_2 \in \mathcal{X}$ such that

$$\|S_{\theta^*}(x_1) - S_{\theta^*}(x_2)\| \leq \varepsilon$$

we have, since R_φ is (X, M, ε_0) -simple, that

$$\|\hat{x}_{1\varphi} - \hat{x}_{2\varphi}\| = \|R_\varphi(S_{\theta^*}(x_1)) - R_\varphi(S_{\theta^*}(x_2))\| \leq l_R \cdot \varepsilon$$

Thus:

$$\|x_1 - x_2\| \leq \|x_1 - \hat{x}_{1\varphi}\| + \|\hat{x}_{1\varphi} - \hat{x}_{2\varphi}\| + \|x_2 - \hat{x}_{2\varphi}\| \leq 2\sqrt{L} + l_R \cdot \varepsilon$$

We now show that every $\varepsilon \leq \varepsilon_0$ holds $2\sqrt{L} + l_R \cdot \varepsilon < 2\sqrt{\bar{L}}$. First, note that

$$\varepsilon_0 = \frac{\frac{\sqrt{2}-1}{2} \sqrt{\text{Var}[X]}}{l_R} = \frac{(1 - \frac{1}{\sqrt{2}}) \sqrt{\frac{1}{2} \text{Var}[X]}}{l_R}$$

In addition, since $L \leq \frac{\bar{L}}{2}$,

$$\sqrt{\bar{L}} - \sqrt{L} \geq \sqrt{\bar{L}} - \sqrt{\frac{\bar{L}}{2}} = (1 - \frac{1}{\sqrt{2}}) \sqrt{\bar{L}}$$

so for any $\varepsilon \leq \varepsilon_0$, we get:

$$\varepsilon \leq \varepsilon_0 = \frac{(1 - \frac{1}{\sqrt{2}}) \sqrt{\bar{L}}}{l_R} < \frac{2(1 - \frac{1}{\sqrt{2}}) \sqrt{\bar{L}}}{l_R} \leq \frac{2(\sqrt{\bar{L}} - \sqrt{L})}{l_R}$$

Which finally leads to

$$2\sqrt{L} + l_R \cdot \varepsilon < 2\sqrt{\bar{L}}$$

Thus, for any $x_1, x_2 \in \mathcal{X}$ that hold $\|S_{\theta^*}(x_1) - S_{\theta^*}(x_2)\| \leq \varepsilon \leq \varepsilon_0$, we get:

$$\begin{aligned}
\|x_1 - x_2\|^2 &\leq (2\sqrt{\bar{L}} + l_R \cdot \varepsilon)^2 \\
&< (2\sqrt{\bar{L}})^2 = 4\bar{L} \\
&= 2 \cdot \mathbb{E}_{x \sim X} \|x - \bar{X}\|^2 \\
&= \mathbb{E}_{x \sim X} \|x - \bar{X}\|^2 + \mathbb{E}_{x \sim X} \|x - \bar{X}\|^2 - 2 \cdot \mathbb{E}_{x \sim X} [x - \bar{X}]^T \mathbb{E}_{x \sim X} [x - \bar{X}] \\
&= \mathbb{E}_{x_1, x_2 \sim X} \|(x_1 - \bar{X}) - (x_2 - \bar{X})\|^2 \\
&= \mathbb{E}_{x_1, x_2 \sim X} \|x_1 - x_2\|^2
\end{aligned}$$

Taking expectation yields the desired result for Definition 4:

$$\mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2 \mid \|S_\theta(x_1) - S_\theta(x_2)\| \leq \varepsilon] < \mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2]$$

□

Theorem 6.2. There exists a discrimination game $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$, $\varepsilon_0 \geq \varepsilon_M$ and a receiver $\varphi \in \Phi$ which is (X, M, ε_0) -simple and non-degenerate, where a synchronized sender matching R_φ is not ε -spatially meaningful for any ε .

Proof of Theorem 6.2. We propose the following counter example. Let $M = \{1, 2, 3, 4, 5, 6\}$ and let X be the uniform distribution over the finite support set $\{1, \dots, 6\} \cup \{-1, \dots, -6\}$. Let $\varepsilon_0 = \varepsilon_M = 1$. We let both hypothesis classes be unrestricted, set the number of candidates d to 2 (a single distractor), and denote the following sets:

$$\begin{aligned}
A_1 &= \{1, -1\} \\
A_2 &= \{2, -2\} \\
A_3 &= \{3, -3\} \\
A_4 &= \{4, -4\} \\
A_5 &= \{5, -5\} \\
A_6 &= \{6, -6\}
\end{aligned}$$

We construct a receiver function that optimally tells apart members from different sets, but is unable to differentiate within each one. Formally:

- R_φ cannot tell apart inputs within each set:

$$\forall k \in \{1, \dots, 6\}, \forall x_1, x_2 \in A_k, \forall m \in M : R(m, x_1, x_2) = (0.5, 0.5)$$

- R_φ is optimal for inputs from different sets; Given the message $m = k$ where $k \in \{1, \dots, 6\}$, R_φ outputs the true probabilities of each input being the target conditioned on the event $X \in A_k$:

$$R_\varphi(k, x_1, x_2) = \begin{cases} (0.5, 0.5) & x_1 \in A_k \text{ and } x_2 \in A_k \\ (1, 0) & x_1 \in A_k \text{ and } x_2 \notin A_k \\ (0, 1) & x_1 \notin A_k \text{ and } x_2 \in A_k \end{cases}$$

The largest possible distance between R_φ 's outputs is $\frac{1}{\sqrt{2}}$, and the minimal distance between messages is $\varepsilon_M = 1$. Thus,

$$\|R_\varphi(x_1) - R_\varphi(x_2)\| \leq \frac{1}{\sqrt{2}} \cdot 1 \leq \frac{1}{\sqrt{2}} \cdot \|x_1 - x_2\|.$$

Note that

$$\frac{\sqrt{2}-1}{2\varepsilon_0} \cdot \sqrt{\text{Var}[X]} = \frac{\sqrt{2}-1}{2} \cdot \sqrt{\frac{91}{6}} > \frac{1}{\sqrt{2}}$$

Which means R_φ is (X, M, ε_0) -simple. We now show that the sender that maps each input to the index of its set, i.e.,

$$S_{\tilde{\theta}}(i) = \begin{cases} 1 & \text{if } x \in A_1 \\ 2 & \text{if } x \in A_2 \\ 3 & \text{if } x \in A_3 \\ 4 & \text{if } x \in A_4 \\ 5 & \text{if } x \in A_5 \\ 6 & \text{if } x \in A_6 \end{cases}$$

is synchronized for R_φ . Let $x \in \mathcal{X}$ belong to set A_k . The loss function is bounded by:

$$\begin{aligned} \ell(\theta, \varphi, x) &= \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim X}} \left[-\log \begin{cases} R_\varphi(S_\theta(i), x, x_2)_1 & \text{if } t = 1 \\ R_\varphi(S_\theta(i), x_2, x)_2 & \text{if } t = 2 \end{cases} \right] \\ &= P(X \in A_k) \cdot \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim X|_{A_k}}} \left[-\log \begin{cases} R_\varphi(S_\theta(x), x, x_2)_1 & \text{if } t = 1 \\ R_\varphi(S_\theta(x), x_2, x)_2 & \text{if } t = 2 \end{cases} \right] \\ &\quad + P(X \notin A_k) \cdot \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim X|_{X \setminus A_k}}} \left[-\log \begin{cases} R_\varphi(S_\theta(x), x, x_2)_1 & \text{if } t = 1 \\ R_\varphi(S_\theta(x), x_2, x)_2 & \text{if } t = 2 \end{cases} \right] \\ &= P(X \in A_k) \cdot \left(-\log \frac{1}{2} \right) + P(X \notin A_k) \cdot \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim X|_{X \setminus A_k}}} \left[-\log \begin{cases} R_\varphi(S_\theta(i), x, x_2)_1 & \text{if } t = 1 \\ R_\varphi(S_\theta(i), x_2, x)_2 & \text{if } t = 2 \end{cases} \right] \\ &\geq P(X \in A_k) \cdot (\log 2) + P(X \notin A_k) \cdot (-\log 1) \\ &= \ell(\tilde{\theta}, \varphi, x). \end{aligned}$$

This means that $\tilde{\theta}$ is synchronized for R_φ . In fact, $S_{\tilde{\theta}}$ is optimal by Corollary 5.4. We can continue simplifying the expression to show that R_φ is strictly better than constant:

$$\begin{aligned} \ell(\tilde{\theta}, \varphi, x) &= P(X \in A_k) \cdot (\log 2) + P(X \notin A_k) \cdot (-\log 1) \\ &= \frac{1}{6} \cdot \log 2 \\ &< \frac{1}{4} \cdot \log 2 \end{aligned}$$

And indeed, the average loss of the optimal constant is:

$$\begin{aligned} \ell(\theta, \varphi^C, i) &= \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim I}} \left[-\log \begin{cases} R_{\varphi^C}(S_\theta(i), x, x_2)_1 & \text{if } t = 1 \\ R_{\varphi^C}(S_\theta(i), x_2, x)_2 & \text{if } t = 2 \end{cases} \right] \\ &= \mathbb{E}_{\substack{t \sim U\{1,2\} \\ x_2 \sim X}} \left[-\log \begin{cases} C & \text{if } t = 1 \\ 1 - C & \text{if } t = 2 \end{cases} \right] \\ &= \frac{1}{2} \cdot (-\log C) + \frac{1}{2} \cdot (-\log(1 - C)) \\ &= -\frac{1}{2} \cdot (\log c + \log(1 - c)) \\ \nabla \ell(\theta, \varphi^C, x) &= -\frac{1}{2} \cdot \left(\frac{1}{C} - \frac{1}{1 - C} \right) \Rightarrow C^* = \frac{1}{2} \\ \mathbb{E}_x \ell(\theta, \varphi^{C^*}, x) &= -\frac{1}{2} \cdot 2 \cdot \log \frac{1}{2} = \log 2 \end{aligned}$$

Finally, we can show that $S_{\tilde{\theta}}$ does not induce a spatially meaningful latent space. We make the observation that

$$\mathbb{E}[X \mid S_{\tilde{\theta}}(X) = m] = 0 \quad \forall m \in \text{Img}(S_{\tilde{\theta}})$$

This is a constant function in m which means $\text{Var}(\mathbb{E}[X \mid S_{\tilde{\theta}}(X) = m]) = 0$. By variance decomposition:

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}(\text{Var}[X \mid S_{\tilde{\theta}}(X) = m]) + \text{Var}(\mathbb{E}[X \mid S_{\tilde{\theta}}(X) = m]) \\ \text{Var}[X] &= \mathbb{E}(\text{Var}[X \mid S_{\tilde{\theta}}(X) = m]) \\ \mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2] &= \mathbb{E}_{m} \mathbb{E}_{x_1, x_2 \sim X_{[m]}_{\tilde{\theta}}} [\|x_1 - x_2\|^2] \\ \mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2] &= \mathbb{E}_{x_1, x_2 \sim X} [\|x_1 - x_2\|^2 \mid S_{\theta}(x_1) = S_{\theta}(x_2)]\end{aligned}$$

This equality contradicts Definition 4 that requires strict inequality (for any $\varepsilon_0 > 0$, the requirement will not hold with $\varepsilon = \min(\varepsilon_0, 0.9)$.), showing that $S_{\tilde{\theta}}$ is neither meaning consistent nor spatially meaningful. \square

C.4 Proofs for variations of discrimination setups

Lemma A.1. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a global discrimination game. Assuming Φ is unrestricted, a sender S_{θ} is optimal if and only if it minimizes the following objective:

$$-I(X; S_{\theta}(X))$$

Proof of Lemma A.1. Since negative log-likelihood is a proper scoring rule and Φ is unrestricted, the synchronized receiver for sender S is

$$R^*(m) = f_X(\cdot \mid S(X) = m)$$

Which means that S^* is optimal if and only if:

$$\begin{aligned}S^* &\in \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \ell(S, R^*, x) \\ &= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} -\log f_X(x \mid S(X) = S(x)) \\ &= \arg \min_{S \in \Theta} H(X \mid S(X)) \\ &= \arg \min_{S \in \Theta} -I(X; S(X))\end{aligned}$$

\square

Lemma A.2. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a 2-candidates supervised discrimination game. Assuming Φ is unrestricted, a sender S_{θ} is optimal if and only if it minimizes the following objective:

$$\sum_{m \in M} P(S_{\theta}(X) = m)^2 - \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(S_{\theta}(X) = m, \text{label}(X) = y)^2$$

Proof of Lemma A.2. Since negative log-likelihood is a proper scoring rule (shown in the proof of Lemma 5.3) and Φ is unrestricted, the synchronized receiver for sender S is

$$R^*(m, x_1, x_2) = \mathbb{P}(t \mid x_1, x_2, m)$$

By the game construction, we have

$$\begin{aligned}P(t = j, x_1, x_2, m) &= f_X(x_j) \cdot f_X(x_{3-j} \mid Y \neq \text{label}(x_j)) \cdot \frac{1}{2} \cdot \mathbf{1}\{S(x_j) = m\} \\ &= \frac{f_X(x_1)f_X(x_2)}{P(Y \neq \text{label}(x_j))} \cdot \frac{1}{2} \cdot \mathbf{1}\{S(x_j) = m\}\end{aligned}$$

Since Y is assumed to have uniform distribution, $P(Y \neq \text{label}(x_j)) = \frac{|\mathcal{Y}| - 1}{|\mathcal{Y}|}$, which means

$$P(t = j, x_1, x_2, m) = f_X(x_1)f_X(x_2) \cdot \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} \cdot \frac{1}{2} \cdot \mathbf{1}\{S(x_j) = m\}$$

Thus, the synchronized receiver is

$$\begin{aligned}
P(R^*(m, x_1, x_2) = j) &= P(t = j \mid x_1, x_2, m) \\
&= \frac{P(t = j, x_1, x_2, m)}{P(t = j, x_1, x_2, m) + P(t = 3 - j, x_1, x_2, m)} \\
&= \frac{\mathbf{1}\{S(x_j) = m\}}{\mathbf{1}\{S(x_j) = m\} + \mathbf{1}\{S(x_{3-j}) = m\}}
\end{aligned}$$

Applying this synchronized receiver to the loss function, we get that S^* is optimal if and only if:

$$\begin{aligned}
S^* &\in \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \ell(S, R^*, x) \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \mathbb{E}_{t \sim U\{1,2\}} - \log \left\{ \begin{array}{ll} P(R^*(S(x), x, x') = 1), & \text{if } t = 1 \\ P(R^*(S(x), x', x) = 2), & \text{if } t = 2 \end{array} \right\} \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \mathbb{E}_{x' \sim P(X|Y \neq \text{label}(x))} - \log \frac{1}{1 + \mathbf{1}\{S(x') = S(x)\}} \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \mathbb{E}_{x' \sim P(X|Y \neq \text{label}(x))} \log(1 + \mathbf{1}\{S(x') = S(x)\}) \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} P(X \notin [S(x)] \mid Y \neq \text{label}(x)) \cdot \log 1 + P(X \in [S(x)] \mid Y \neq \text{label}(x)) \cdot \log 2 \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} P(X \in [S(x)] \mid Y \neq \text{label}(x)) \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y) \cdot \mathbb{E}_{x \sim X} P(X \in [S(x)] \mid Y \neq \text{label}(x), S(X) = m, Y = y) \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y) \cdot P(X \in [m] \mid Y \neq y) \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y) \cdot P(X \in [m], Y \neq y) \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y) \cdot (P(X \in [m]) - P(X \in [m], Y = y)) \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m]) \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y) - \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y)^2 \\
&= \arg \min_{S \in \Theta} \sum_{m \in M} P(X \in [m])^2 - \sum_{m \in M} \sum_{y \in \mathcal{Y}} P(X \in [m], Y = y)^2
\end{aligned}$$

□

Lemma A.3. Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a classification discrimination game. Assuming Φ is unrestricted, a sender S_θ is optimal if and only if it minimizes the following objective:

$$-I(Y; S_\theta(X))$$

Proof of Lemma A.3. Since negative log-likelihood is a proper scoring rule (shown in the proof of Lemma 5.3) and Φ is unrestricted, the synchronized receiver for sender S is

$$R^*(m, x_1, x_2) = \mathbb{P}(t \mid x_1, \dots, x_n, m)$$

By the game construction, we have

$$P(t = j, x_1, \dots, x_n, m) = P(X \in [m], Y = j) \cdot \prod_{i=1}^n f_X(x_i \mid Y = i)$$

Thus, the synchronized receiver is

$$\begin{aligned}
P(R^*(m, x_1, \dots, x_n) = j) &= P(t = j \mid x_1, \dots, x_n, m) \\
&= \frac{P(X \in [m], Y = j)}{P(X \in [m])} \\
&= P(Y = j \mid X \in [m])
\end{aligned}$$

Applying this synchronized receiver to the loss function, we get that S^* is optimal if and only if:

$$\begin{aligned}
S^* &\in \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \ell(S, R^*, x) \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} \mathbb{E}_{\{x_i \sim P(X|Y=i)\}_{i=1}^n} - \log P(R^*(S(x), x_1, \dots, x_n) = \text{label}(x)) \\
&= \arg \min_{S \in \Theta} \mathbb{E}_{x \sim X} - \log P(Y = \text{label}(x) | X \in [S(x)]) \\
&= \arg \min_{S \in \Theta} H(Y|S(X)) \\
&= \arg \min_{S \in \Theta} -I(Y; S(X))
\end{aligned}$$

□

D Details on empirical work

D.1 Training details

Data. We use a version of the Shapes dataset [22], where each input is an image of dimensions $(3, 64, 64)$ containing a single object. The object has a random shape out of (square, rectangle, triangle, pentagon, cross, circle, semicircle, ellipse), and a random color out of (red, green, blue, yellow, magenta, cyan, gray). In addition, the object has a random position, rotation, size and distortion. See the Shapes github repo and our code for exact details. We generate $\sim 10K$ data samples, split into 9088 (142 batches of 64) training samples and 896 (14 batches of 64) validation samples. The MNIST dataset [30] contains $(1, 28, 28)$ images of hand-written digits, split into $\sim 54K$ training samples and 5440 validation samples.

Agent architectures and loss functions. We use the format {Shapes value}/{MNIST value} to indicate values that differ between the datasets. The sender consists of two parts: an *image encoder* and a *message generator*. The image encoder receives an input image of dimensions $(3, 64, 64)/(1, 28, 28)$ and outputs a latent vector of dimension 100. The exact encoder architecture is given in Figure 5a / Figure 6a. A trainable linear layer then maps this output to a vector of dimension 150 which serves as input to the message generator, which is implemented by a single-layer GRU with hidden dimension 150. Its input is used as an initial hidden state, together with a learnable BOS vector. After the GRU cell, the resulting hidden state is mapped by a linear layer to the vocabulary size (10). We use the Gumbel-Softmax relaxation to approximate a symbol sampling, then an embedding matrix is used on the relaxed on-hot vector and the result is fed to the next generation step (along with the current hidden state).

The receiver has different architectures for reconstruction and discrimination. Both receivers start with a *message embedder*, implemented by a single-layer GRU with hidden size 150. In the reconstruction setup, the final hidden state from that network is given to an *image decoder* implemented by transpose-convolutions, see Figure 5b / Figure 6b for full details. The loss in the reconstruction setting is simply the Mean Squared Error (MSE) between the receiver’s output and the target image. In the discrimination setting, the receiver uses another image encoder (with the architecture from Figure 5a / Figure 6a) and applies it to every image in the set of candidates. The resulting vector is further processed by a Multilayer perceptron (MLP) with two layers, hidden dimension 150 and relu activation. The output of this MLP is a representation for each candidate. We take a dot product of this representation with the final hidden state of the message embedder (both normalized) to get the score for each candidate. The discrimination loss is then the Cross-Entropy classification loss with the target position as label (also called the infoNCE objective). Visual illustrations of the two tasks are given in Figure 7/Figure 8.

Training procedure. While the architectures described above can be jointly trained from scratch, the discrete optimization with Gumbel-Softmax is much slower and less stable than continuous training. We account for that by employing a two-stage training procedure. In the first stage, we train a continuous autoencoder, which is a concatenation of an image encoder and an image decoder. The trained image encoder is used by the sender agent in both setups, and also by the receiver in the discrimination game. The trained decoder is used by the receiver in the reconstruction game. Both of these pretrained modules are kept frozen during the second stage of training, in which the discrete

game is played out. This means that during EC training, the sender only optimizes the message generator (and a linear layer), and the receiver only optimizes the message embedder (and the MLP in the discrimination setting).

Training parameters. We train (during the EC phase) for 200/100 epochs with batch size 64. We use vocabulary size 10 and message length 4. Every message is full length as we do not use a special end-of-sentence token. We use a fixed temperature value of 1.0 for the Gumbel-Softmax sampling method.

Computation resources. The main training stage (EC) takes ~ 4 hours on Shapes, ~ 1.5 hours on MNIST using NVIDIA GeForce RTX 2080 Ti. The continuous training takes < 10 minutes.

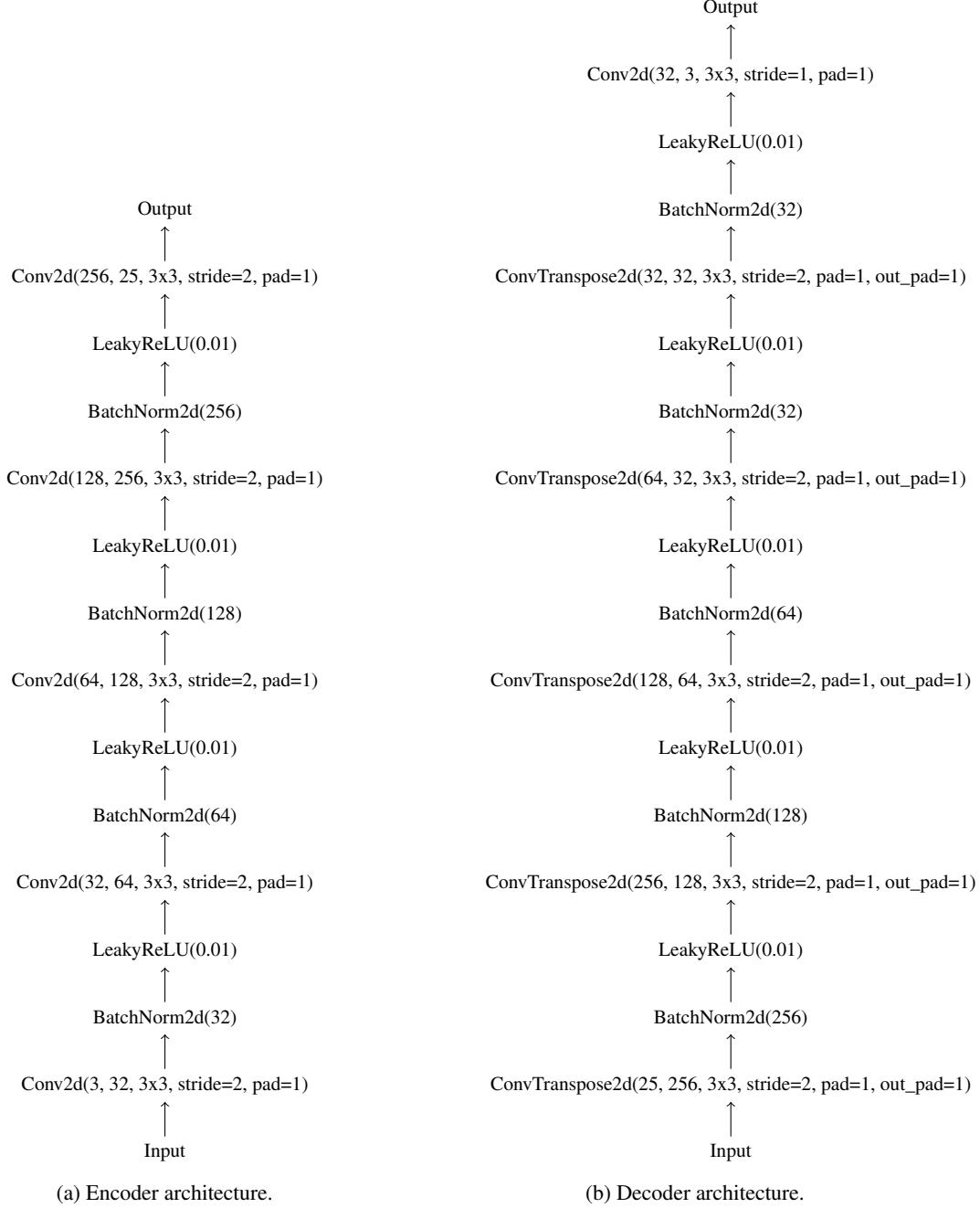


Figure 5: Encoder and Decoder architectures used on Shapes.

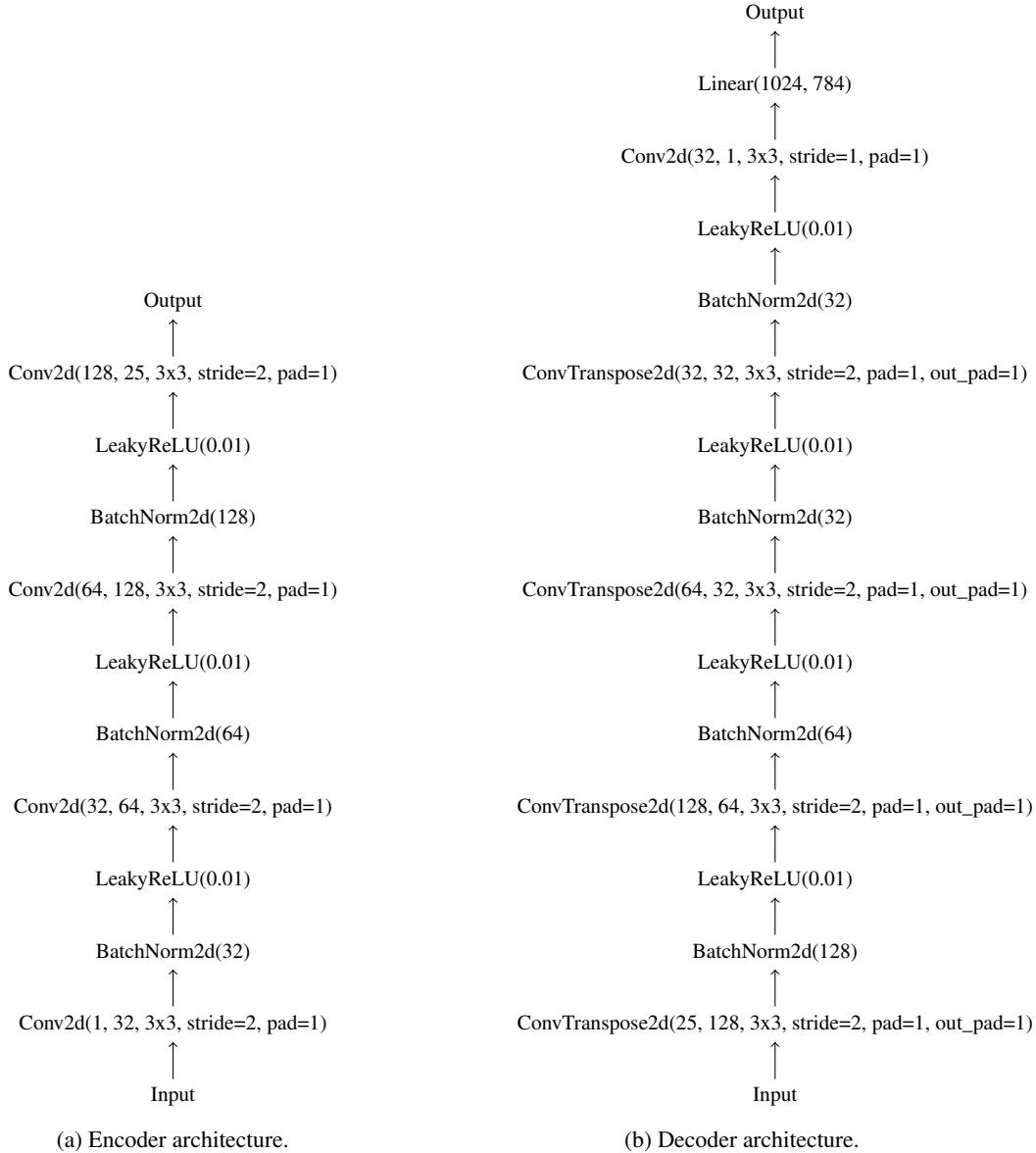


Figure 6: Encoder and Decoder architectures used on MNIST.

D.2 Metrics

Discrimination accuracy. We sample 40 distractors independently, so the receiver predicts the target out of 41 given candidates. Every image in the validation set is used as target exactly once, and the reported accuracy is the percentage of images that were correctly predicted as targets. In the reconstruction setting, the prediction is defined by first passing the given message to the receiver, which outputs a reconstruction of the target, and then picking the candidate which minimizes the loss function (Euclidean distance) with respect to that output.

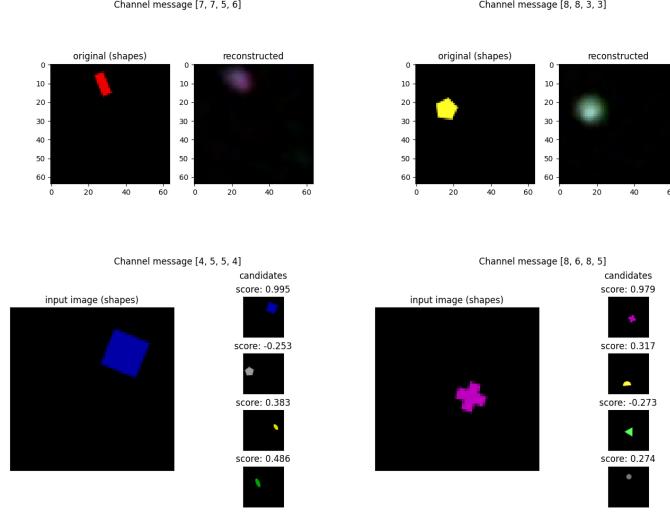


Figure 7: Reconstruction (top) and discrimination (bottom) examples on Shapes

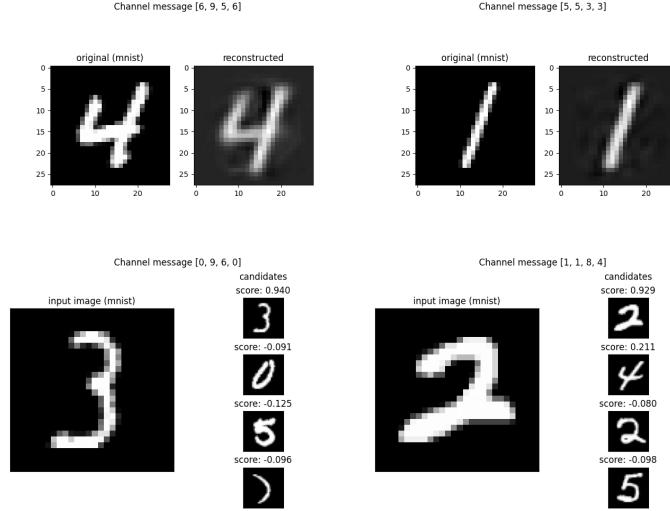


Figure 8: Reconstruction (top) and discrimination (bottom) examples on MNIST

Message variance. The unexplained variance is $\mathbb{E} [\text{Var}[X \mid S(X)]] = \sum_{m \in M} P(m) \cdot \text{Var} [[m]]$ where $[m]$ is the set of images mapped to message m . Using the empirical measures

$$\begin{aligned}\text{Var} [[m]] &= \frac{1}{2 \cdot |[m]|^2} \sum_{x_1, x_2 \in [m]} \|x_1 - x_2\|^2 \\ P(m) &= \frac{|[m]|}{N}, \quad N = 896\end{aligned}$$

we get

$$\begin{aligned}\sum_{m \in M} P(m) \cdot \text{Var} [[m]] &= \sum_{m \in M} \frac{|[m]|}{2 \cdot |[m]|^2 \cdot N} \sum_{x_1, x_2 \in [m]} \|x_1 - x_2\|^2 \\ &= \frac{1}{2N} \sum_{m \in M} \frac{1}{|[m]|} \sum_{x_1, x_2 \in [m]} \|x_1 - x_2\|^2\end{aligned}$$

We calculate message variance using this formula. A lower value indicates more semantic consistency (less unexplained variance). See Algorithm 1 for the full calculation procedure. Instead of calculating

Euclidean distance between raw pixels [4, 10], which might be uninformative, we use embeddings created by the continuous encoder (perceptual loss [39]).

Algorithm 1 Compute message variance

Require: Dataset X , Set of possible messages M , Sender function $S : X \rightarrow M$

- 1: Initialize dictionary $equiv_classes$ with keys from M and empty lists as values
- 2: **for** each input $x \in X$ **do**
- 3: $message \leftarrow S(x)$
- 4: Append x to $equiv_classes[message]$
- 5: **end for**
- 6: Initialize $pairwise_sum \leftarrow 0$
- 7: **for** each $message \in M$ **do**
- 8: Initialize $local_sum \leftarrow 0$
- 9: **for** pair (x_i, x_j) in $equiv_classes[message]$ **do**
- 10: $local_sum \leftarrow local_sum + \text{distance}(x_i, x_j)$
- 11: **end for**
- 12: **if** $\text{len}(equiv_classes[message]) > 1$ **then**
- 13: $local_sum \leftarrow local_sum / \text{len}(equiv_classes[message])$
- 14: $pairwise_sum \leftarrow pairwise_sum + local_sum$
- 15: **end if**
- 16: **end for**
- 17: $N \leftarrow \text{size of } X$
- 18: $result \leftarrow pairwise_sum / (2 \times N)$
- 19: **return** $result$

Random baseline. The distribution of messages induced by the trained sender significantly affects some of the metrics. To facilitate an insightful comparison between the two EC setups, we establish a random baseline per-protocol which randomizes the assignment of images while preserving the number of inputs mapped to each message. Comparing the outcomes across different baselines allows us to examine the extent to which channel utilization impacts the results.

Topographic similarity. TopSim [5] is defined as the (Spearman) correlation between pairwise distances in the input space and corresponding distances in the message space. We use edit distance in the message space, and Euclidean distance between embeddings in the input space. The embeddings are generated by the frozen encoder trained in the continuous stage (see Appendix D.1).

Bag-of-symbols disentanglement. BosDis [8] calculates per-symbol the gap between the highest and second highest mutual information to an attribute, normalizes by the symbol’s entropy and sums over the vocabulary. We consider only two attributes (shape and color), so the mutual information gap is simply the difference between the two.

Positional disentanglement. PosDis [8] operates similarly to BosDis, but considers positions within the message rather than symbols.

Speaker-centered topographic similarity. S-TopSim [13] operates similarly to PosDis, but rather than summing over positions, the mutual information gap is calculated per attribute. This is a more straightforward application of MIG [9].

Message purity. The purity of a message with respect to an attribute is the percentage of images in its equivalence class that have the majority value. The minimal purity value is achieved by a uniform distribution over all of the attribute’s possible values; the maximal value, 100%, is reached only if all images mapped to the message share the same value (e.g., if the attribute is shape and all images mapped to that message contain squares). We take a weighted average of message purity over messages. The final measure can be interpreted as the highest attainable accuracy by a deterministic classifier of messages to the attribute. An easier version of this measure, which we term max-purity, defines the purity of each message as the maximum purity over all attributes.

E Additional experiments

E.1 Full results: Compositionality, semantic consistency and game performance

We consider three additional compositionality measures from literature (BosDis, PosDis and S-PosDis), see Appendix D.2 for their definitions. Some of these measures require multiple attributes (based on the mutual information gap), so we report results on Shapes. Table 3 shows these results, and Table 4 shows the correlation between each pair of measures, over three subsets of the collected measurements: the joint set (10 runs), reconstruction only (5 runs) and discrimination only (5 runs). We note that the correlations calculated over the joint set are heavily affected by the two distributions and often show opposite values to the individual tables. Furthermore, the results regarding S-PosDis are somewhat noisy due to its low values reported in Table 3. A key observation is that the correlation between compositionality and game performance (disc. accuracy) is negative or nonexistent across all compositionality measures in both the reconstruction subset and the discrimination subset, other than the S-PosDis value in the discrimination table (which we believe is noise, see previous note). This further corroborates our findings in Section 7.

Table 3: Empirical results with compositionality measures from literature.

	TopSim ↑	BosDis ↑	PosDis ↑	S-PosDis ↑	Disc. accuracy ↑	Msg Variance ↓
Reconstruction	0.34 ±0.02	0.11 ±0.01	0.04 ±0.03	0.01 ±0.01	0.32 ±0.03	1334.38 ±78.05
Discrimination	0.09 ±0.01	0.32 ±0.19	0.19 ±0.13	0.01 ±0.00	0.62 ±0.05	2280.24 ±157.38

E.2 Message purity over the Shapes dataset

Figure 9 presents the color purity, shape purity and max purity in each game type. See Appendix D.2 for their definitions. Table 5 shows the baseline values as well. The reconstruction-trained agents generate messages that convey more information about each attribute. However, the generated distribution of messages (which defines the random baseline) seems to have significant effect on the purity results. We hypothesize that higher multiplicity of attributes may better differentiate the two tasks.

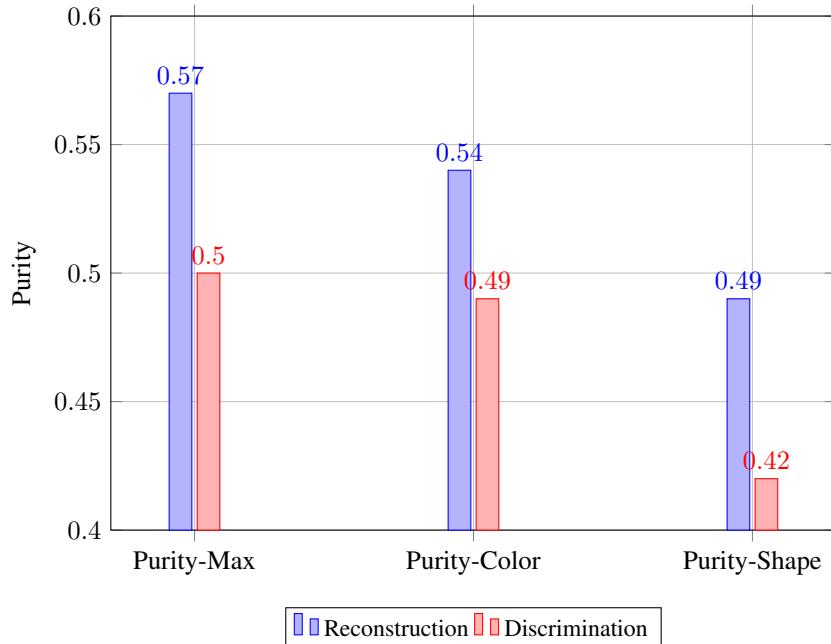


Figure 9: Message purity per attribute and game type.

Table 4: Correlation Matrices for All Data, Reconstruction, and Discrimination Groups

(a) Correlation Matrix (All Data)

	TopSim	BosDis	PosDis	S-PosDis	Accuracy	Msg Variance
TopSim	1.00	-0.63	-0.63	0.49	-0.97	-0.97
BosDis	-0.63	1.00	0.98	-0.26	0.65	0.61
PosDis	-0.63	0.98	1.00	-0.29	0.61	0.62
S-PosDis	0.49	-0.26	-0.29	1.00	-0.40	-0.42
Accuracy	-0.97	0.65	0.61	-0.40	1.00	0.92
Msg Variance	-0.97	0.61	0.62	-0.42	0.92	1.00

(b) Correlation Matrix (Reconstruction)

	TopSim	BosDis	PosDis	S-PosDis	Accuracy	Msg Variance
TopSim	1.00	0.33	-0.37	0.76	-0.24	-0.61
BosDis	0.33	1.00	0.00	0.17	-0.37	-0.93
PosDis	-0.37	0.00	1.00	-0.02	-0.61	0.26
S-PosDis	0.76	0.17	-0.02	1.00	-0.03	-0.28
Accuracy	-0.24	-0.37	-0.61	-0.03	1.00	0.35
Msg Variance	-0.61	-0.93	0.26	-0.28	0.35	1.00

(c) Correlation Matrix (Discrimination)

	TopSim	BosDis	PosDis	S-PosDis	Accuracy	Msg Variance
TopSim	1.00	0.61	0.50	0.82	0.10	0.00
BosDis	0.61	1.00	0.99	0.07	-0.04	-0.23
PosDis	0.50	0.99	1.00	-0.07	-0.14	-0.17
S-PosDis	0.82	0.07	-0.07	1.00	0.26	0.12
Accuracy	0.10	-0.04	-0.14	0.26	1.00	-0.80
Msg Variance	0.00	-0.23	-0.17	0.12	-0.80	1.00

Table 5: Purity measures compared with random baselines.

	Purity-Max		Purity-Color		Purity-Shape	
	Trained	Random	Trained	Random	Trained	Random
Reconstruction	0.57 ± 0.03	0.54 ± 0.02	0.54 ± 0.03	0.49 ± 0.02	0.49 ± 0.02	0.48 ± 0.03
Discrimination	0.50 ± 0.05	0.45 ± 0.03	0.49 ± 0.05	0.43 ± 0.04	0.42 ± 0.03	0.42 ± 0.03

E.3 Spatial meaningfulness empirical analysis

Recall that spatial meaningfulness requires close messages to have similar meanings. We propose an evaluation method that alters the message space such that it is well clustered, and evaluates the semantic consistency of message clusters rather than individual messages. To do so, we partition the 10 vocabulary symbols to five groups: $\{0, 1\}$, $\{2, 3\}$, $\{4, 5\}$, $\{6, 7\}$, and $\{8, 9\}$. During training, agents are restricted to using messages composed of symbols from a single group. This is implemented by letting the sender generate the first symbol freely, and then masking out all symbols except those belonging to the same set as the initial symbol for the remainder of the message generation. Consequently, messages such as $(3, 2, 3, 2)$ and $(7, 7, 7, 7)$ are possible, whereas a message like $(3, 4, 5, 9)$ can never be generated. There are 16 possible messages within each cluster, and only messages from the same cluster can have a Hamming distance smaller than 4 (the maximum). We then calculate spatial meaningfulness as the empirical unexplained variance (Algorithm 1) of these

clusters; a lower value indicates more spatially meaningful messages. Table 6 and Table 7 present this measurement over the validation sets of Shapes and MNIST respectively. While the reconstruction setting leads to a better improvement over the random baseline on Shapes, as expected by our findings, the results do not demonstrate high levels of spatial meaningfulness in any of the trained protocols. We hypothesize that the size of each cluster (16 messages) is too large, and perhaps a different partitioning method will provide more insight.

Table 6: Cluster Variance on Shapes.

EC setup	Trained Cluster Var ↓	Random Cluster Var	Improvement
Reconstruction	3742.19 ± 64.87	3915.19 ± 70.52	4.41%
Discrimination	3804.09 ± 84.73	3919.11 ± 73.73	2.94%

Table 7: Cluster Variance on MNIST.

EC setup	Trained Cluster Var ↓	Random Cluster Var	Improvement
Reconstruction	1790.88 ± 29.88	1921.78 ± 0.17	6.81%
Discrimination	1745.86 ± 29.47	1922.23 ± 0.43	9.18%
Supervised disc.	1783.60 ± 44.15	1922.06 ± 0.35	7.20%

F Alternative formulations of the discrimination game

Recall that we formally model the discrimination receiver as a function $R: M \times \mathcal{X}^d \rightarrow \Delta\{1, \dots, d\}$ (Definition 2). In this setting, assuming an unrestricted hypothesis class, we show that the conditionally optimal receiver outputs a uniform distribution over the subset of candidates that belongs to the given message’s equivalence class. This functionality is demonstrated in Figure 10.

In this section, we tackle two alternative formulations which provide stronger resemblance to common implementations of the discrimination game, at the cost of added complexity. The first alleviates the receiver’s ability to consider more than one candidate at a time; the second explicitly models the prediction using dot product in a latent space. With the assumption of an unrestricted receiver hypothesis class, our results are persistent.

$$\begin{aligned}
 S(\text{candidate1}) &= m \\
 S(\text{candidate2}) &\neq m \\
 S(\text{candidate3}) &\neq m \quad \xrightarrow{\text{---}} R^*(m, \text{candidates}) = \begin{cases} \text{Candidate1} & \text{w.p. } 0.5 \\ \text{Candidate2} & \text{w.p. } 0 \\ \text{Candidate3} & \text{w.p. } 0 \\ \text{Candidate4} & \text{w.p. } 0.5 \\ \text{Candidate5} & \text{w.p. } 0 \end{cases} \\
 S(\text{candidate4}) &= m \\
 S(\text{candidate5}) &\neq m
 \end{aligned}$$

Figure 10: The conditionally optimal discrimination receiver, assuming it is available in Φ .

Candidate-unaware discrimination

In this formulation, rather than directly outputting a distribution over the candidates, the receiver predicts a score for each of the candidates independently, and those scores are normalized to create the final probabilities. Referred to as the descriptive-only variation by [12]. Formally, the receiver takes the form

$$R: M \times \mathcal{X} \rightarrow [0, 1]$$

and the game objective is

$$\ell(S, R, x) := \mathbb{E}_{\substack{t \sim U\{1, \dots, d\}, x_t=x \\ x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d \sim X^{d-1}}} - \log \frac{R(S(x_t), x_t)}{\sum_{i=1}^d R(S(x_t), x_i)}.$$

The key difference is that each score only depends on the corresponding candidate and the message. However, the conditionally optimal receiver from Figure 10 can still be achieved, for example by the following receiver:

$$R^*(m, x) = \begin{cases} 1 & \text{if } S(x) = m \\ 0 & \text{otherwise} \end{cases}.$$

By assuming an unrestricted Φ , this receiver is available and therefore optimal. Hence, the set of optimal communication protocols is the same in this setting as in the original one.

Latent similarity discrimination

This formulation further restricts the receiver by generating the scores via cosine similarity in a latent space, and using the softmax operator to create the final probabilities. This formulation closely resembles common implementations of the discrimination game. Formally, let $z \in \mathbb{N}$ be the latent dimension, the receiver now consists of two parts: an input encoder $R_X : \mathcal{X} \rightarrow \mathbb{R}^z$ and a message encoder $R_M : M \rightarrow \mathbb{R}^z$. The game objective becomes

$$\ell(S, R, x) := \mathbb{E}_{\substack{t \sim U\{1, \dots, d\}, x_t = x \\ x_1, \dots, x_{(t-1)}, x_{(t+1)}, \dots, x_d \sim X^{d-1}}} - \log \frac{\exp(\cos(R_M(S(x_t)), R_X(x_t)))}{\sum_{i=1}^d \exp(\cos(R_M(S(x_t)), R_X(x_i)))}.$$

Note that the softmax output is strictly positive in every entry, so the conditionally optimal receiver from Figure 10 cannot be achieved in this alternative formulation. However, we make the observation that the optimal input encoder outputs the same latent vector as the corresponding message encoder:

$$\begin{aligned} & - \log \frac{\exp(\cos(R_M(S(x_t)), R_X(x_t)))}{\sum_{i=1}^d \exp(\cos(R_M(S(x_t)), R_X(x_i)))} \\ &= - \log \left(1 - \frac{\sum_{i \neq t} \exp(\cos(R_M(S(x_t)), R_X(x_i)))}{\exp(\cos(R_M(S(x_t)), R_X(x_t))) + \sum_{i \neq t} \exp(\cos(R_M(S(x_t)), R_X(x_i)))} \right) \\ &\geq - \log \left(1 - \frac{\sum_{i \neq t} \exp(\cos(R_M(S(x_t)), R_X(x_i)))}{e + \sum_{i \neq t} \exp(\cos(R_M(S(x_t)), R_X(x_i)))} \right) \end{aligned}$$

and this lower bound can be achieved by selecting

$$R_X^*(x) = R_M(S(x)).$$

In other words, the conditionally optimal receiver needs to map equivalent inputs to the same latent vector (up to a multiplicative constant). In the case of unrestricted receiver hypothesis class, this still allows the relationship between equivalent inputs to be arbitrary. However, the representations of different messages are incentivized to be well separated, meaning that if similar inputs are mapped to different messages, their encodings by R_X^* will be far apart. Restrictions on R_X^* may render this infeasible, opening the way for semantic consistency in the discrimination setting. We leave further investigation of this idea to future work.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main results are outlined in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Discussion of this work's limitation is given in Section 8.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Assumptions are properly defined and stated for each theorem. Complete proofs are given in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We give a rigorous explanation of the training setup and procedure in Appendix D.1, and release code with reproducibility instructions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Exact commands for reproducibility are given with the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Training details, including data splits and hyperparameters, are given in Appendix D.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the mean and standard deviation in our results, see Section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Compute information is given in Appendix D.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research presented in the paper adheres to the NeurIPS Code of Ethics in all aspects.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This theoretical work may improve interpretability of learned representations in the long run, but has no direct societal impacts that we foresee.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This is a theoretical work and does not include released data or models, so there is no risk of this paper's contributions being misused.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the most relevant code package (EGG) in Section 7.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Training details are given in Appendix D.1. The code contains documentation and usage instructions for reproducibility.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No experiments in this paper involve human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No experiments in this paper involve human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.