





1 We sincerely thank all reviewers for their feedback. We present an image reference game where it is necessary to model
 2 other agents’ understanding of task-related concepts to succeed. The reviewers indicate that our framework is shown
 3 through a technically sound experimental evaluation (R1) to be capable of modeling other agents’ expertise (R2,R3),
 4 has relevance to human interaction domains (R1,R2), and is well positioned to inspire future research in settings with
 5 dynamic agent behaviors (R2). We will clarify the issues raised below and incorporate them into our final version.

6 **R3: Training V and active policy.** V is trained by minimizing MSE on all practice and evaluation games independent
 7 of the speaker variant. Attributes in evaluation games are chosen greedily using V, in practice games different
 8 parameterized policies are used. Active policy explicitly minimizes the MSE for V. Optimizing directly for high reward,
 9 i.e. $R = \sum_k r_k$, does not necessarily increase information content for understanding the listener; e.g. active and epsilon
 10 greedy policies obtain equivalent downstream performance (Fig2) but active policy achieves lower VI (Fig4).

11 **R2: Reactive baseline, define N + M (L144).** This is a static policy which always uses the same attribute until a
 12 negative reward is encountered (L167), at which point a new attribute is sampled. “Sequence of episodes” refers to
 13 all of the $N + M$ games played with each listener, respectively defined as the number of practice games the speaker
 14 uses to understand the listener, and the number of evaluation games to verify the speaker’s model. Reactive baseline
 15 remembers utilized attributes when going from practice to evaluation games, disregarding attributes that didn’t work.

16 **R1, R2: Qualitative example.** As an example, we train a speaker with 5 listener
 17 populations that respectively do not understand color, shape, size, length, and
 18 pattern attributes. Due to the lack of space, we show two games from a randomly
 19 sampled test set sequence of the trained speaker policy with a color blind agent.
 20 We will extend this with more examples in the main paper.

Maximally					
Game	discrim.	Chosen	Target	Confounder	Why?
6	Green breast	Green breast			Listener is colorblind
76	Yellow leg	Upland ground like shape			

21 **R2, R3: Related work.** Our work is indeed applicable to robotics: robots could
 22 reason about users’ understanding of object properties when describing object locations [C], as well as about multiple
 23 other agents’ intentions via a probabilistic generative model, which could extend our model to cooperative tasks [B].
 24 Modeling users’ understanding of an AI’s mind [A] could provide an explanation component to teach users about the
 25 AI. We would like to disentangle two orthogonal aspects of communication, i.e. modeling of other agents and language
 26 learning. While [D,E,F] focus on language learning ([D,E] present synthetic reference games and [D,F] use two agents),
 27 we focus on agent modeling on real-world images on a population of agents as they may have different capacities to
 28 understand the task-related concepts. Our model with emergent language is an interesting extension.

29 **R2: Speaker’s mental model of listener human-interpretable?** The clusters formed with agent embeddings (L194-
 30 201) correlate well with the ground truth clustering of the population, since all speakers with agent embeddings gradually
 31 minimize the VI score as more games are played (Fig4). Moreover, the performance increase with agent embeddings
 32 in Fig3 suggests that the learned function V encodes listeners’ conceptual understanding. Since the attributes are
 33 inherently interpretable, V’s output is a human-interpretable representation of the speaker’s model of the listener.

34 **R2: Without agent embeddings.** After each game, the speaker incorporates the outcome into the agent embedding
 35 (using an LSTM, Fig1), allowing it to form a model of the listener; the embedding is used to condition the attribute
 36 selection policy, i.e. Description Generation Module. With “no agent embedding”, a zero-vector carrying no information
 37 between games is used. Hence, the speaker must maximize performance without storing information about the listener.

38 **R2: Error bars in Fig2, R3: sampling of target and confounder.** Each experiment was run with 3 seeds (random
 39 initializations). Error bars represent the standard deviation. Dataset splits, i.e. training and test images, remain the same
 40 across seeds. Target and confounder images are randomly sampled from the training set (during training) or test set (at
 41 test time). Practice and evaluation games occur both during training (with parameter updates) and testing (no updates).

42 **R2: Variation of Information metric.** $VI(C, C') = H(C) + H(C') - 2I(C, C')$ measures how much information
 43 is lost or gained by switching from a clustering C to C’ where H is the entropy of a clustering and I is the mutual
 44 information between two clusters. The lower VI the better, entailing a close correlation between clusters.

45 **R2: Shared perception module.** As in reality, confusion between different agents may occur also due to how they
 46 perceive their environment. To maintain generality, we allow our framework to have separate perception modules.

47 **R2: Is the information content or the message lost? (L84)** The information content itself does not suffer due to the
 48 listener’s misunderstanding, since the speaker communicates losslessly. The listener’s ability to use the information in
 49 the message suffers, since it is difficult for the listener to properly compare images using a poorly understood attribute.

50 **R1,R2: Writing and Code.** We will remove “Ties to RL” (L77); release code, data and models upon acceptance.

51 [A] Chandrasekaran et al., It Takes Two to Tango: Towards Theory of AI’s Mind, CVPR 2017 [B] Butterfield et al., Modeling
 52 Aspects of Theory of Mind with Markov Random Fields, IJSR 2009 [C] Warnier et al., When the robot puts itself in your shoes.
 53 Managing and exploiting human and robot beliefs, IEEE RO-MAN 2012 [D] Kottur et al., Natural Language Does Not Emerge
 54 ‘Naturally’ in Multi-Agent Dialog, EMNLP 2017 [E] Cogswell et al., Emergence of Compositional Language with Deep Generational
 55 Transmission, arXiv 2019 [F] Das, et al., Learning cooperative visual dialog agents with deep reinforcement learning, CVPR 2017