Learning to Mediate Disparities Towards Pragmatic Communication

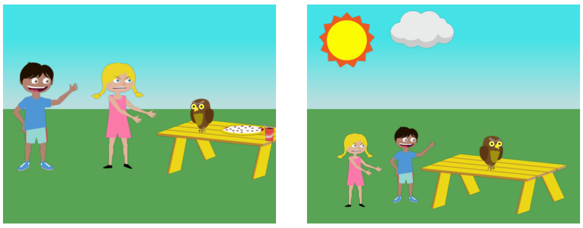
Yuwei Bao[†] **Sayan Ghosh**[§*] **Joyce Chai**[†]
[†]Computer Science and Engineering, University of Michigan [§]ASAPP
{yuweibao, sayghosh, chaijy}@umich.edu

Abstract

Human communication is a collaborative process. Speakers, on top of conveying their own intent, adjust the content and language expressions by taking the listeners into account, including their knowledge background, personalities, and physical capabilities. Towards building AI agents with similar abilities in language communication, we propose Pragmatic Rational Speaker (PRS), a framework extending Rational Speech Act (RSA). The PRS attempts to learn the speaker-listener disparity and adjust the speech accordingly, by adding a light-weighted disparity adjustment layer into *working memory* on top of speaker's *long-term* memory system. By fixing the long-term memory, the PRS only needs to update its working memory to learn and adapt to different types of listeners. To validate our framework, we create a dataset that simulates different types of speaker-listener disparities in the context of referential games. Our empirical results demonstrate that the PRS is able to shift its output towards the language that listeners are able to understand, significantly improve the collaborative task outcome.

1 Introduction

In human communication, speakers often adjust their language production by taking into consideration listeners' personality, background knowledge, perceptual or physical capabilities etc (Clark, 1996). Recent years have seen an increasing amount of work that explores pragmatic reasoning based on Rational Speech Act (RSA) (Andreas and Klein, 2016; Fried et al., 2018a,b; White et al., 2020; Cohn-Gordon et al., 2018), multi-agent emergent communication framework (Lazaridou et al., 2020; Lazaridou and Baroni, 2020), and Theory of Mind in communication (Bara et al., 2021; Zhu et al., 2021). However, except for (Zhu et al., 2021), most previous works assume that the listeners and the speakers have the same background knowledge and capabilities, including vocabulary size, visual access, and relative locations. This assumption is a great simplification of real-world communication where speakers and listeners often have various types of disparities.



(a) Target (b) Distractor

.

Literal Speaker: There is an owl on the table.
Rational Speaker: There is a pizza on the table.
Listener's Disparity: understands hypernym of food only.
Pragmatic Rational Speaker: There is food on the table.

Figure 1: **TASK:** Given two images, the speaker generates a description for the target image and asks the listener to pick out the image described. Both players win if the listener picks the correct one. In this example, a Literal Speaker could generate multiple captions that suit the target, such as the one above, whereas a Rational Speaker limits the description to the unique features of the target (e.g. pizza). If the listener only understands the hypernym of food (disparity), a Pragmatic Rational Speaker would learn the disparity and use *food* instead of *pizza* to help the listener understand.

To address this limitation, this paper extends the Rational Speech Act (RSA) (Frank and Goodman, 2012) model towards rational agents learning to adapt behaviors based on their experience with the listener. The design choice of our model is inspired by the human cognitive system (Cowan, 2008; Wardlow, 2013) where a limited capacity ***working memory*** is built on top of the ***long-term memory*** to adjust the output to be task and environment specific. Each communication is a modification on the long-term memory (Reed, 2012) with situation-specific factors. In our framework, we fix the long-term memory which captures lan-

---

[*] Work done during undergraduate study at the University of Michigan.

guage structure for communication, and introduce a light-weighted working memory (Miyake and Shah, 1999) for the Pragmatic Rational Speaker to modify and accommodate two goals: 1) a **task goal** which retrieves relevant information from the long-term memory and accomplish the task, and 2) a **disparity goal** which learns and adjusts the conversation to accommodate the listener's disparity through reinforcement learning. We separate each component as they are independent of each other in utility, and can be easily switched and adapted for new tasks and new environment.

Different from previous works which only demonstrate how learned models affect task performance (e.g. (Shridhar et al., 2020; Zhu et al., 2021; Corona et al., 2019)), one of our goals is to also provide transparency on what models have indeed learned towards the end goal. It's well established that end-to-end neural models can often take advantage of spurious data bias to gain end performance. Models that only report end measure without showing their internal works would not be sufficient to tell the whole story about model's abilities.

To serve this goal, we situated our investigation in the context of a referential game[1] as shown in Figure 1. We carefully curated a dataset to simulate two types of disparity: *knowledge disparity* and *perceptual disparity*. Our empirical results demonstrate that our model is able to significantly improve the collaborative game performance by shifting communication towards the language that the listeners with disparities are able to understand. In addition, our results show that separating working memory from long-term memory leads to faster learning and better performance than the previous model which conducted joint end-to-end learning.

Our contributions are the following. 1) Following human cognition, we demonstrate the benefits of separating working memory from the long-term memory, compared to end-to-end joint training. 2) We propose a new dataset to simulate multiple distinct types of disparities, and demonstrate the pragmatic adaptability of our model. 3) Instead of focusing on mere end task performance, we show model's strong language shift ability to accommodate listener's disparities. The dataset and code are available through `https://github.com/sled-group/Pragmatic-Rational-Speaker` to facilitate future work on pragmatics and theory of mind in language interpretation and generation.

2 Related Work

It has been studied (Leung et al., 2021; Stephens et al., 2010; Wardlow, 2013) in psychology that human speakers adjust the way how we speak for successful communication after learning the listener's disparity. Some recent work (Zarrieß and Schlangen, 2019; Zhu et al., 2021; Corona et al., 2019; Hawkins et al., 2021) attempt to address similar questions. We build our model upon the following two concepts.

Rational Speech Act (RSA)

The Rational Speech Act (RSA) model (Frank and Goodman, 2012) is a probabilistic model for the speakers and listeners to pragmatically reason about each other's intention. In the context of a referential game (Monroe and Potts, 2015), for example (Figure 1), given an image m, it starts with a literal speaker S_0 to generate caption c: $P_{S_0}(c|m)$. A rational listener L_1 reasons about the literal speaker's (S_0) strategy and picks the best image that matches the description. A rational speaker S_1 then takes the rational listener's (L_1) strategy into account and produces a caption c that maximizes the collaborative game goal.

$$P_{L_1}(m|c) \propto P_{S_0}(c|m) \cdot P(m)$$

$$P_{S_1}(c|m) \propto P_{L_1}(m|c) \cdot P(c)$$

In previous work (Andreas and Klein, 2016) and (Lazaridou et al., 2020; Lazaridou and Baroni, 2020), the same referential game setup was used to propose a rational speaker that learns to reason the collaborative game and to produce natural sounding image captions based on RSA. However, they were mainly addressing the *task goal*, assuming the speaker and listener have the exact same capabilities and knowledge background, which is unrealistic. In our work, we created listeners with disparity d and extend this model for the speaker to accommodate both the *task* and *disparities* goals.

Working Memory

Working memory (also short-term memory) is used in neuropsychology and cognitive science (Cowan,

---

[1] Different from traditional referential ground work as (Liu et al., 2013; Gorniak and Roy, 2004; Siebert and Schlangen, 2008; DeVault et al., 2005; Liu et al., 2012), we adopted this term from a recent line of work (Lazaridou et al., 2020; Andreas and Klein, 2016) to refer to the task described in Figure 1.

	Role	Long Term	Work Mem - Task	Work Mem - Disparity
S_0	Literal Speaker	Image Caption		
S_1	Rational Speaker	Image Caption	Simulated Listener	
S_1^d	Pragmatic Rational Speaker	Image Caption	Simulated Listener	Disparity Adjustment
L_1	Rational Listener	Caption Grounding		
L_1^d	Rational Listener w/ Disparity	Caption Grounding with disparity		

Table 1: Types of Speaker and Listener

2008; Miyake and Shah, 1999) to refer to the memory that controls attention, plans and carries out behavior. It is a combination of multiple components, including the contribution of long-term memory (Reed, 2012; Sawangjit et al., 2018) and situation-specific task processing (Funahashi, 2017).

The classical artificial intelligence work such as ACT (Heise and Westermann, 1989) and SOAR (Laird et al., 1987) also incorporated the concept of working memory to model human short-term memory. The similar concept has been used in recent work such as (Hermann et al., 2017; Hill et al., 2017). Our work is a novel application of the working memory to pragmatically adjust communication for speaker-listener disparities (*disparity* goal), and take advantage of the internal simulation architecture to achieve the *task* goal.

Similar to (Kottur et al., 2017; Lazaridou et al., 2020), our model learns to converge language to adapt to listener's disparities through interactions, instead of ground truth supervision on language generation. The speakers have zero prior knowledge on the listener's background nor an oracle access to probe the listener's brain.

Different from previous works, our model is able to generalize to distinct types of disparities. In addition, while previous models were trained in an end-to-end joint fashion, our work separates training and demonstrates the efficiency of working memory. Most importantly, few of the previous work were able to showcase model's language capabilities and only evaluate them by the end performance (e.g. accuracy), whereas our work emphasizes on evaluating how well the models learn to shift the language towards better understanding.

3 Dataset

There are many levels of disparities during verbal communication (Stephens et al., 2010), including phonetic, lexical, grammatical, semantic representations, etc. In our work, we assembled two datasets, and challenge the speaker model to handle two types of disparities: 1) knowledge disparity, and 2) perceptual disparity.

The *knowledge disparity* is simulated through the *hypernym* dataset, where the listener only understands the hypernym for all the objects (e.g. "food" instead of "pizza"), whereas the speaker understands both. This dataset challenges the speaker model at the *lexical* level to learn what listener's vocab limitation, and shift towards the words that they understand.

The *perceptual disparity* is simulated through the *limited visual* dataset, where the listener has impaired vision or some objects were physically blocked from the eyesight. This dataset challenges the speaker to shift attention and pick the visible objects for the listener to describe. For control and demonstration purposes, we remove all the animal-related objects and words from listener's training.

These datasets are used to simulate listener's disparities and train the listener's model as described in Section 4.2. The speaker's long term memory was trained with the original data which has full knowledge of the vocab and objects, but no idea what the listeners are or aren't capable of. Detailed dataset components can be found in the Appendix.

We modified the Abstract Scenes (Gilberto Mateos Ortiz et al., 2015) dataset for our experiments. There are 10020 images, each including 3 ground truth captions, and a median of 6 to 7 objects. We assembled ~35k pairs of images that differ by ≤ 4 objects as the *Hard* set, ~25k pairs that differ by > 4 objects as the *Easy* set, and together as the *Combined* set. The image pairs were split into training, validation and testing by a ratio of 8:1:1.

4 Method

Given a pair of images m_0, m_1, the target image indicator $t \in \{0, 1\}$, and the listener's disparity d, the speaker generates a caption c for the target image m_t, and the listener needs to pick out the correct target t given c. Both receive a reward of $+1$ upon correct choice, and -1 otherwise.

Following the RSA model, as shown Figure 2,

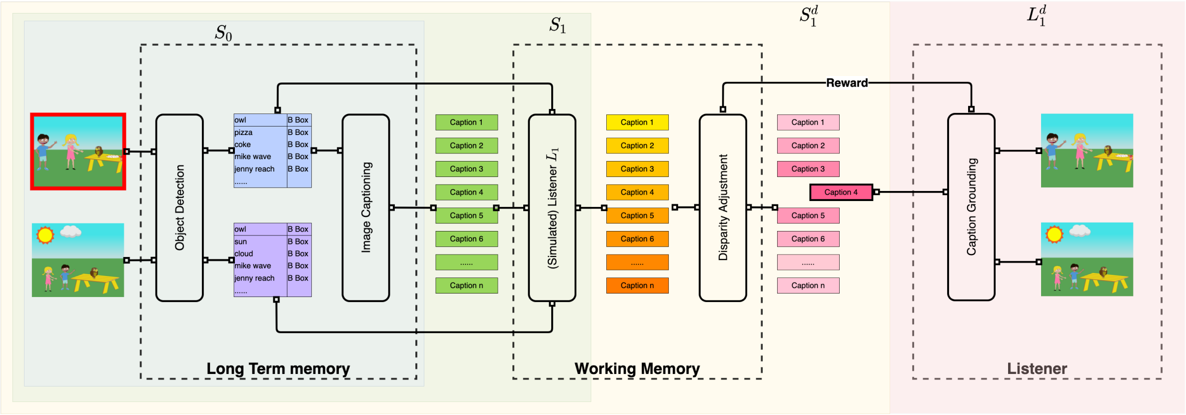


Figure 2: Speaker and Listener Models: Literal Speaker S_0 uses and object detector and image captioning module to generate a list of candidate captions in the fixed long term memory. The Rational Speaker S_1 simulates an internal listener to rank (illustrated by color gradient) the candidate captions by their uniqueness in describing the target image. The Pragmatic Rational Speaker S_1^d interacts with the actual listener to rerank the captions and pick out the best one to accommodate the disparity and the task goal. Both simulated listener and disparity adjustment components are inside the working memory for task specific and disparity specific adjustments.

we start by building the Literal Speaker $\mathbf{S_0}$, gradually increase model structure and functionality with the vanilla Rational Speaker $\mathbf{S_1}$ and the Pragmatic Rational Speaker $\mathbf{S_1^d}$. Upon retrieving a list of candidate captions C from the long-term memory, the final goal for $\mathbf{S_1^d}$ is to output the best caption c in the working memory, that accommodates both 1) **task goal:** describes the unique features of the target image, and 2) **disparity goal:** learns and accommodates the listener's disparity.

Table 1 is a brief summary of each model. The Literal Speaker $\mathbf{S_0}$ generates candidate captions c for a given image m (Eq 1), which serves as the long-term memory. The Rational Listener $\mathbf{L_1}$ picks out an image as the target given speaker's description (Eq 2). The vanilla Rational Speaker $\mathbf{S_1}$ achieves the *task goal* by simulating the listener's mind internally in its working memory (Eq 3). $\mathbf{L_1^d}$ incorporates disparity to the Rational Listener. The Pragmatic Rational Speaker $\mathbf{S_1^d}$ adds a light-weight disparity adjustment layer (Eq 5) to learn and accommodate listener's disparity through interactions, and achieves both goals. Each component can be easily switched and adapted to new tasks or environment.

$$\mathbf{S_0} : P(c|m_t) \tag{1}$$

$$\mathbf{L_1} : P(t|m_0, m_1, c) \propto P_{S_0}(c|m_t) \cdot P(m_t) \tag{2}$$

$$\begin{aligned}\mathbf{S_1} :&P(c|m_0, m_1, t) \propto \\ &P_{L_1}(t|m_0, m_1, c) \cdot P(c|m_0, m_1)\end{aligned} \tag{3}$$

$$\begin{aligned}\mathbf{L_1^d} :&P(t|m_0, m_1, c, d) \propto \\ &P_{S_1}(c|m_0, m_1, t, d) \cdot P(t|m_0, m_1, d)\end{aligned} \tag{4}$$

$$\begin{aligned}\mathbf{S_1^d} :&P(c|m_0, m_1, t, d) \propto \\ &P_{L_1^d}(t|m_0, m_1, c, d) \cdot P(c|m_0, m_1, d)\end{aligned} \tag{5}$$

4.1 Literal Speaker S_0

The Literal Speaker S_0 (Figure 2) is an object detection based image captioning module that generates caption candidates for the target image.

$$\begin{aligned} o_1, \ldots, o_k, b_1, \ldots, b_k &= \texttt{ObjDet}(m_t) \\ e_1, \ldots, e_k &= \texttt{WordEmb}(o_1, \ldots, o_k) \\ c_1, \ldots, c_n &= \texttt{Transformer}(e_1, \ldots, b_1, \ldots) \end{aligned} \tag{6}$$

For a given target image m_t, since it's important to ground words to the scenes in order to control the disparities in vocabularies, we applied the object detector YOLO3 (Redmon and Farhadi, 2018) to extract a list of k detected objects $O = \{o_1, o_2, \ldots, o_k\}$, and their corresponding bounding boxes $B = \{b_1, b_2, \ldots, b_k\}$. Each image chooses at most $max\_obj = 9$ detected objects, and the names of each were embedded with a pre-trained BERT (Devlin et al., 2019) word embedding $E = \{e_1, e_2, \ldots, e_k\}$. These embeddings are then concatenated with their bounding box locations, and sent to the Transformer Decoder to generate $beam\_size = 30$ candidate captions $C = \{c_1, c_2, \ldots, c_n\}$ for each target image.

4.2 Rational Listener (L_1)

Without disparity concerns, the Rational Listener picks out the image that they believe is the target.

$$\begin{aligned} g_0 &= \texttt{FT\_Transformer}(m_0, c) \\ g_1 &= \texttt{FT\_Transformer}(m_1, c) \\ t &= \text{argmax}_{i \in \{0,1\}} \texttt{CosSim}(g_i, c) \end{aligned} \quad (7)$$

Recall that S_0 used a `Transformer` decoder to connect the image and its corresponding captions. We reuse the same Fixed pre-trained Training-mode Transformer module (named `FT_Transformer`) to decide which image does the caption ground better in. Adopting the idea of teacher-forcing language training, the output (g_i) of `FT_Transformer` with an input pair (m_i, c) should closely resemble the original input c if the input image m_i is indeed the one used to generate the caption c. By calculating the cosine similarity of each (g_i, c) pair, the image that grounds better (higher `CosSim`) in the description would be chosen as the target.

This module allows the agents to quickly and accurately make the decisions without further training. In theory, if the speaker and the listener were to have the exact same brain (same model and weights), the performance of this task should approach 100%. The results of "No Disparity" speaker in Figure 3 confirmed the design choice.

4.3 Rational Speaker (S_1)

Without disparity concerns, the Rational Speaker (S_1) fulfills the **task goal** by simulating (Figure 2) the Rational Listener (L_1)'s behavior, and rank the candidate captions generated by the Literal Speaker (S_0) according to how well they can describe the target image apart from the distractors. This design is under the fair assumption that both speakers and listeners are aware of the collaborative game goal, but can be switched for other task purposes.

$$\begin{aligned} &\text{For } i \in \{0, \cdots, n\}, \text{where } n = |C| : \\ &t_i, p_i = \texttt{Simulate\_L}_1(m_0, m_1, c_i) \\ &c = c_{\text{argmax}_i [[t_i == t^*]] \cdot p_i} \end{aligned} \quad (8)$$

Given an image pair (m_0, m_1), and a list of candidate captions $C = \{c_1, \cdots, c_n\}$ generated by S_0, the Rational Speaker goes through each caption c_i and simulates how well the listener (`Simulate_L`$_1$) would pick out the correct target image. If a candidate caption c_i helps the simulator pick out the correct target image (i.e. $t_i == t^*$) with high confidence (p_i), then it will be chosen as the final caption sent over to the actual listener. The simulated listener shares the same architecture as L_1 and initializes the weights pre-trained from S_0. By doing so, the Rational Speaker takes the listener's intention into account and achieves the task goal.

4.4 Listener with Disparities (L_1^d)

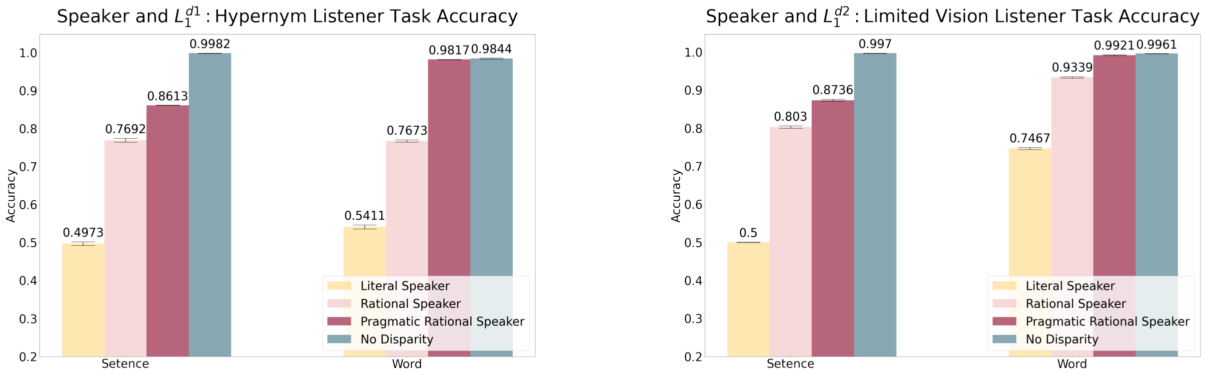
In the real world, however, it is hardly the case that different agents have the exact same knowledge background, experiences, physical capabilities, etc. The listener's decision making process is influenced by various kinds of disparities d.

To study speaker's ability of situated language adjustment, we created two representative types of listeners with different knowledge background and visual capabilities by training different caption grounding modules (`FT_Transformer`) with the datasets assembled in Section 3. These disparities would challenge the speaker model to adjust the language at different levels.

1. L_1^{d1} : Hypernym. With limited vocabulary and knowledge in a certain domain, people tend to refer to objects in their hypernym form (e.g. "animal" instead of "cat"). In this experiment, we create listeners that would refer to all the detected objects by their hypernyms. This disparity would require the speaker to switch individual words that share similar meanings.

2. L_1^{d2} : Limited Visual. Due to the physical orientation or impaired vision capability, it is likely that some objects are blocked or hardly visible to one party but not the other. In this experiment, we remove all the animal objects from listener's visual detected object list (O), and replace the relevant descriptions with the special token '[UNK]'. This disparity would require the speaker to shift attention, and choose alternative objects to describe.

We investigate in listeners with a subset of speaker's capabilities under the argument that in the opposite case, the listener could use only a subset of the knowledge to achieve best performance without having the speakers to adjust the speech. Other disparities can be inferred through transfer learning or are left for further investigation with broader information access and datasets.



(a) L_1^{d1}: Hypernym

(b) L_1^{d2}: Limited Vision

Figure 3: Referential game Accuracy: The Pragmatic Rational Speakers are able to significantly outperform Literal Speakers and vanilla Rational Speakers across different types of disparities. Word level models achieve higher performance and is much closer to the No Disparity upper bound than the sentence level communication.

4.5 Pragmatic Rational Speaker (S_1^d)

On top of the Rational Speaker (S_1), the Pragmatic Rational Speaker incorporates a disparity adjustment layer to learn and accommodate the listener's disparity through emergent communication.

$$\begin{aligned} &\text{For } i \in \{0,\cdots,n\}, \text{where } n=|C|: \\ &q_i = \mathtt{MLP}(\mathtt{SentenceEmb}(c_i)) \\ &a_i = [[t_i == t^*]]\cdot p_i \cdot q_i \\ &c = c_{\text{argmax}_i a_i} \end{aligned} \tag{9}$$

We use a pretrained BERT model to embed each candidate caption c_i, add a single MLP layer, and approximate the REINFORCE policy through Equation 9. The reward (r_{c^*}) for each chosen caption c^* is $+1$ or -1. The loss is calculated for all the chosen captions across each batch (Eq 10).

$$L = -\sum_{c^*} \log(a_{c^*}) \cdot r_{c^*} \tag{10}$$

4.6 Communication with Words

We conducted the same sets of experiments using individual words (object names) instead of sentences to demonstrate the effects of working memory on disparity accommodation and internal task simulation, reducing the noise that came from the imperfection of the image description generator. The simplified pipeline uses the detected object name embedding for disparity adjustment, and the listener picks the target images by conducting simple word matching.

5 Results and Analysis

We evaluate our models (S_0, S_1, S_1^d) on the referential game (Figure 1) along four dimensions: **End-task Performance, Efficiency, Transparency, and Balance of Goals.** Recall that each speaker model has different capabilities (Table 1) and only S_1^d is able to fulfill both *task* and *disparity* goals. Implementation details and more experiment results can be found in Appendix.

1. **[Task Performance]** that measures overall accuracy of the collaborative game. Task performance is often the sole evaluation metrics in previous work.

2. **[Efficiency]** that measures time used for model training across tasks.

3. **[Transparency]** that uncovers the underlying distribution shift of vocabulary use learned to accommodate different types of disparities.

4. **[Balance of Goals]** that the working memory needs to consider between the task and disparity goals to achieve maximum performance

5.1 Task Performance Comparison

To assess the performance of the speakers in the collaborative game, Figure 3 presents the task accuracies with Literal Speaker (S_0), Rational Speaker (S_1), Pragmatic Rational Speaker (S_1^d), and No Disparity (S_1^{nd}). S_1^{nd} has the same structure as S_1 and was trained on the same disparity dataset as the corresponding listener. It serves as the upper bound of performance. The same experiments also were conducted at the word level.

For each type of listener disparity, the performance is $S_0 << S_1 < S_1^d < S_1^{nd}$. The vanilla Rational Speaker (S_1) improved the overall performance from Literal Speaker by over 25% because it is achieving the task goal to describe the target image apart from the distractor. The Pragmatic Rational Speaker (S_1^d) is able to learn and adjust for the listener's disparity, and further improve the game performance by $\sim$10%. There is still, however a gap between S_1^d and the upper bound S_1^{nd}, where the speaker and the listener have the exact knowledge and capability limitation, potentially due to the imperfection in caption generations.

Breaking down between the *hard*, *easy* datasets in Figure 4 (recall that image pairs that differ by ≤ 4 objects are in the *Hard* set, otherwise the *Easy* set), S_1^d on the *easy* dataset is able to gain a lot more improvement upon its Rational Speaker compared to the pair trained on the *hard* dataset. The gap between S_1^d and No Disparity is also a lot smaller for the model trained on the *easy* dataset. This is likely because when a pair of images differ more objects (easier), the model has more options to adjust upon, hence the larger improvement.

Compared to the sentence level model, the word level pragmatic speaker for L_1^{d1} achieves even higher improvement against the corresponding Rational Speaker. They both achieve almost perfect accuracy with close to zero gap to the upper bound. This suggests the high potential of the disparity adjustment design, especially after reducing the caption generation and interpretation noise.

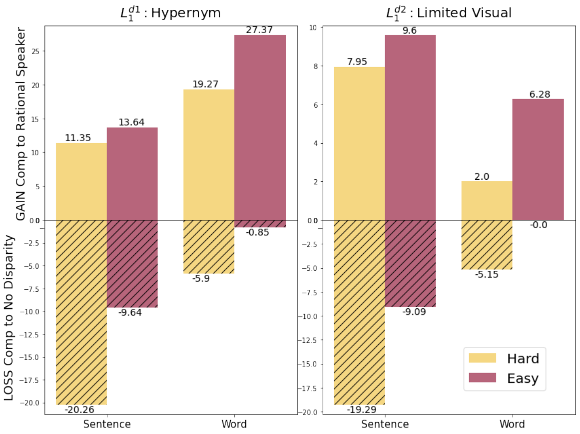


Figure 4: Pragmatic Rational Speaker Performance Gain/Loss compared to Rational Speaker and No Disparity (upper bound).

5.2 Learning Efficiency

To study the training efficiency of the working memory, we compared our model to the joint training "Multi-Task leaning" model in (Lazaridou et al., 2020)'s work, retrained and evaluated in our dataset. The image captioning model and the REINFORCE learning are joint trained through a combined loss:

$$L = \lambda_f L^{functional} + \lambda_s L^{structural}$$

Functional in our task refers to the REINFORCE learning to achieve both task and disparity goals (evaluated by Accuracy), and *structural* refers to the caption generation loss for natural-sounding language (evaluated by BLEU4). We used $\lambda_f = \lambda_s = 1$ as in previous work for our experiments.

	Train(min)	Accuracy%	BLEU4
Joint	19.04	60.14	27.79
Separate	21.02	**77.34**	**29.3**
• LM	11.59		29.3
• WM	**9.43**	77.34	

(a) L_1^{d1}: Hypernym

	Train(min)	Accuracy%	BLEU4
Joint	29.52	63.69	27.29
Separate	29.95	**81.09**	**29.3**
• LM	11.59		29.3
• WM	**18.36**	81.09	

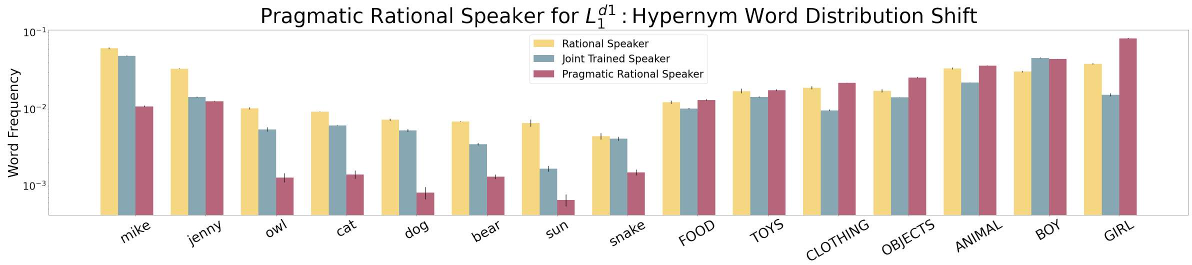
(b) L_1^{d2}: Limited Vision

Table 2: Compared to joint training, separate training only needs to train the long-term memory once, and can achieve higher performance. **LM**: Long-term Memory, **WM**: Working Memory.

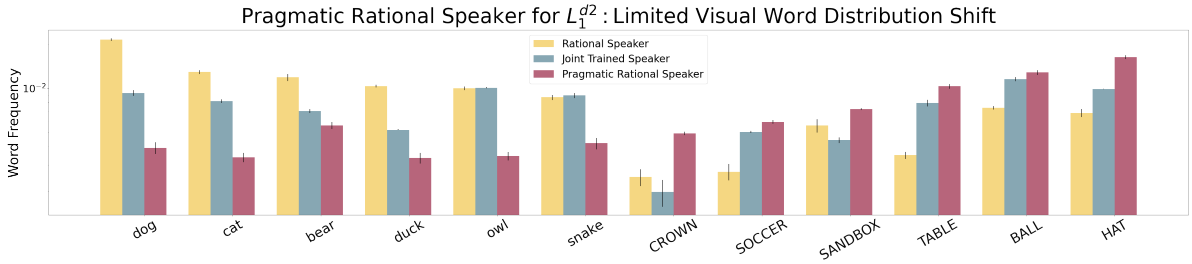
Detailed training and comparison strategies can be found in the Appendix. Table 2 shows that for each type of disparity, our model separating working memory from long-term memory is able to achieve higher accuracy and higher BLEU4 score than the joint training. Moreover, the Joint Trained model needs to retrain all the weights for each type of disparity from scratch, whereas our model only needs to train the long-term memory once, and retrain the light weighted working memory for each type of disparity, which is much more efficient.

5.3 Transparency: Vocabulary Adjustment

To gain insights in whether the Pragmatic Rational Speaker (PRS) is actually adjusting the descriptions for listeners' disparities or taking the advantage of statistical bias to achieve higher task performance, we plotted the word distribution shift across different types of disparities. Qualitative examples can be found in Figure 6. For each experiment, the word frequencies of all the chosen captions were calculated for the Rational Speakers, the Pragmatic Rational Speakers, and Joint Training. We collected the top choice of each speaker per image

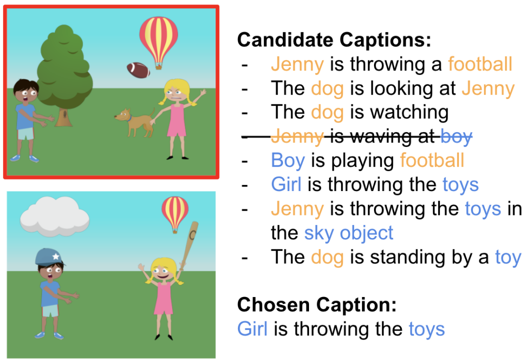


(a) L_1^{d1}: Hypernym



(b) L_1^{d2}: Limited Vision

Figure 5: Word Distribution Shift: The Pragmatic Rational Speaker for L_1^{d1} avoids specific object names and prioritizes some hypernyms. The Pragmatic Rational Speaker for L_1^{d2} avoids animal related words in communication.

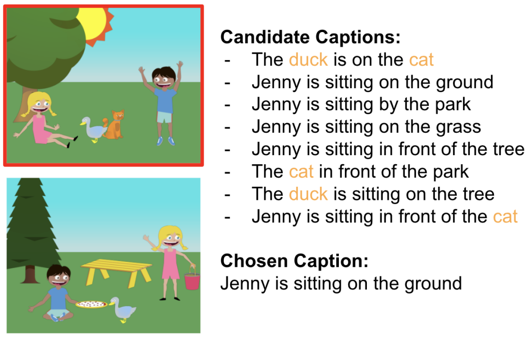


Candidate Captions:

- Jenny is throwing a football
- The dog is looking at Jenny
- The dog is watching
- ~~Jenny is waving at boy~~
- Boy is playing football
- Girl is throwing the toys
- Jenny is throwing the toys in the sky object
- The dog is standing by a toy

Chosen Caption:
Girl is throwing the toys

(a) L_1^{d1}: Hypernym



Candidate Captions:

- The duck is on the cat
- Jenny is sitting on the ground
- Jenny is sitting by the park
- Jenny is sitting on the grass
- Jenny is sitting in front of the tree
- The cat in front of the park
- The duck is sitting on the tree
- Jenny is sitting in front of the cat

Chosen Caption:
Jenny is sitting on the ground

(b) L_1^{d2}: Limited Vision

Figure 6: Qualitative examples for each disparity adjustment. The orange words are the vocabulary that the Pragmatic Rational Speaker is avoiding, and the blue words are the preferred alternative for the listeners. The ~~strikethrough~~ sentences are discarded because they can be used to describe both images.

pair, repeated the experiments 3 times, and reported the mean and standard deviation in Figure 5.

In the Hypernym disparity (Figure 5a) experiment, where the listener only understands the hypernym of detected objects, the lower-case words on the left are the top detected object names, and the upper-case words on the right are hypernyms. On the left side, the word frequencies of PRS significantly dropped from the Rational Speaker. On the right side, the model is maintaining similar level, or using some of the hypernyms more frequently (y-axis in log scale). Note that the Rational Speaker can generate both hypernym and hyponym regardless of disparities, and multiple valid captions available for all speakers to choose from. For the Joint Trained Speaker, we also observed a hyponym usage drop (left), but it's unclear how it accommodates the disparity without using hypernyms. This result shows that PRS learned to avoid using hyponyms, and replaced them with their hypernym to accommodate the disparity.

For the Limited Visual disparity (Figure 5b), since all the animal objects are missing for the listener, there is a sharp decline in S_1^{d2}'s use of animal related words during the communication. Instead, it is choosing other objects such as "hat", and "ball" to describe the target image. The PRS is accommodating listener L_1^{d2}'s disparity by shifting the attention and choosing alternative objects other than animals to communicate. The behavior of the Joint Trained Speaker is harder to interpret.

5.4 Balancing Between Goals

Recall that the working memory of the Pragmatic Rational Speaker (S_1^d) has two two goals: 1) **Task Goal**: an internal simulation of a listener to rank the candidate captions by their uniqueness in describing the target image, and 2) **Disparity Goal**: a disparity adjustment layer to learn and accommodate the listener's disparity through interactions. Each goal component can be formalized in the above two terms (Equation 11). We parameterized each term with λ_l and λ_d to study how different $\lambda_l : \lambda_d$ weight ratio could affect rational speaker's ability to achieve both goals.

$$a_i = ([[t_i == t^*]] \cdot p_i)^{\lambda_l} \cdot (q_i)^{\lambda_d} \qquad (11)$$

Figure 7 shows that when the Pragmatic Rational Speaker puts a high emphasis on adjusting the listener's disparity λ_d, it would "forget" to describe the unique characters of the target image and lower the overall performance. On the other hand when the PRS emphasize too much on the task goal, it would "forget" to accommodate listener's disparities, and lower the overall performance as well. In the end, we chose $\lambda_l : \lambda_d = 1 : 1$ for all experiments demonstrated above.

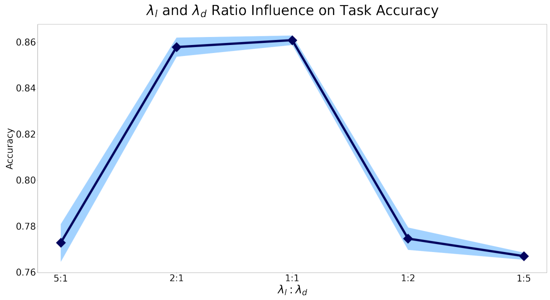


Figure 7: Balancing between the task goal and the disparity adjustment goal: the Pragmatic Rational Speaker needs a balanced emphasise on both λ_l and λ_d in order to achieve both goals simultaneously.

6 Conclusion and Future Work

In this work, we present a novel framework based on the Rational Speech Act framework for pragmatic communication that can adjust the conversation content according to listener's disparities by adding a light-weighted working memory on top of speaker's long-term memory. The Pragmatic Rational Speaker significantly improved the collaborative game performance by shifting the conversation towards the language that the listeners are able to understand. The flexibility and training efficiency also makes it easy to be applied broadly.

There are, however, several limitations that requires further investigation. First of all, despite recent progress, algorithms that connect language and the visual world are still limited. For example, caption generation, even in this simple setup, often does not faithfully capture what's been conveyed in the images. As our framework heavily relies on the quality of various models that bridge language and vision, e.g., as part of our long term memory, it's important to improve functionality and performance of these base models.

We conducted our experiments in a relative simple and artificial environment with the purpose of easy control and demonstration. We emphasize on evaluating model's actual language ability of adjusting for the disparities on top of task performance. The next step would be to apply the framework to more realistic images and interactive environment.

Other than listener's knowledge background and perceptual capabilities, there are a lot of other reasons for language communication to be adjusted, such as the physical environment, relative positions, speaker's personalities, etc. Studying how a rational agent can accommodate these disparities would require additional multimodal datasets and information processing methods.

At the moment, the Pragmatic Rational Speaker trains a new layer in working memory from scratch for each type of disparity. This could have backward influence on the long-term memory. In lifelong learning (Parisi et al., 2019) like humans, the working memory can shape their long-term memory. At the very least, the model could store each learned disparity adjustments for future encounter. This modification is left for future work.

Last but not least, instead of training for every single type of disparity to name, human learners have the ability of meta-learning and zero-shot transferring existing knowledge to a new category. Future work on pragmatic reasoning should be easily adaptable to different disparities and situations.

Acknowledgements

This work was supported in part by the National Science Foundation under grant IIS-1949634. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Rodolfo Corona, Stephan Alaniz, and Zeynep Akata. 2019. Modeling conceptual understanding in image reference games. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13155–13165.

Nelson Cowan. 2008. Chapter 20 what are the differences between long-term, short-term, and working memory? In Wayne S. Sossin, Jean-Claude Lacaille, Vincent F. Castellucci, and Sylvie Belleville, editors, *Essence of Memory*, volume 169 of *Progress in Brain Research*, pages 323–338. Elsevier.

David DeVault, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, page 1–4, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018a. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018b. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3318–3329.

Shintaro Funahashi. 2017. Working memory in the prefrontal cortex. *Brain Sci.*, 7(5):49.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. 2015. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515, Denver, Colorado. Association for Computational Linguistics.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *J. Artif. Int. Res.*, 21(1):429–470.

Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. 2021. From partners to populations: A hierarchical bayesian account of coordination and convention.

Elke Heise and Rainer Westermann. 1989. *Anderson's Theory of Cognitive Architecture (ACT*)*, pages 103–127. Springer Berlin Heidelberg, Berlin, Heidelberg.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech M. Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. 2017. Grounded language learning in a simulated 3d world. *ArXiv*, abs/1706.06551.

Felix Hill, Karl Moritz Hermann, Phil Blunsom, and Stephen Clark. 2017. Understanding grounded language learning agents. *CoRR*, abs/1710.09867.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. 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, Copenhagen, Denmark. Association for Computational Linguistics.

John E. Laird, Allen Newell, and Paul S. Rosenbloom. 1987. Soar: An architecture for general intelligence. *Artif. Intell.*, 33(1):1–64.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *ArXiv*, abs/2006.02419.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Ashley Leung, Alexandra Tunkel, and Daniel Yurovsky. 2021. Parents fine-tune their speech to children's vocabulary knowledge. *Psychological Science*, 32(7):975–984. PMID: 34212788.

Changsong Liu, Rui Fang, and Joyce Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea. Association for Computational Linguistics.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Chai. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France. Association for Computational Linguistics.

Akira Miyake and Priti Shah. 1999. *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge University Press.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *CoRR*, abs/1510.06807.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767.

Stephen K. Reed. 2012. *Human Cognitive Architecture*, pages 1452–1455. Springer US, Boston, MA.

Anuck Sawangjit, Carlos N Oyanedel, Niels Niethard, Carolina Salazar, Jan Born, and Marion Inostroza. 2018. The hippocampus is crucial for forming non-hippocampal long-term memory during sleep. *Nature*, 564(7734):109–113.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *CoRR*, abs/2010.03768.

Alexander Siebert and David Schlangen. 2008. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 84–87, Columbus, Ohio. Association for Computational Linguistics.

Greg J. Stephens, Lauren J. Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32):14425–14430.

Liane Wardlow. 2013. Individual differences in speakers' perspective taking: the roles of executive control and working memory. *Psychon Bull Rev*, 20(4):766–772.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. *ArXiv*, abs/2006.00418.

Sina Zarrieß and David Schlangen. 2019. Know what you don't know: Modeling a pragmatic speaker that refers to objects of unknown categories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 654–659, Florence, Italy. Association for Computational Linguistics.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pages 12901–12911. PMLR.

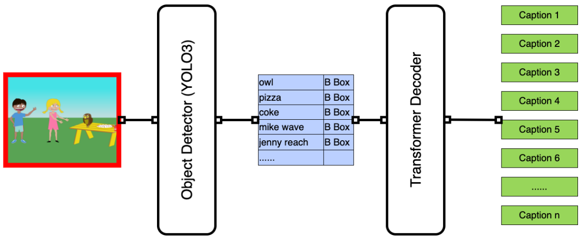
A Speaker and Listener Model Architecture Breakdown

(a) **Literal Speaker** S_0: for each input image, we run YOLO3 object detector to get a list of detected object names. Each name is embedded with pre-trained BERT embedding, and concatenated with their bounding box location. The embedded images goes through a Transformer Decoder to generate a list of candidate captions.

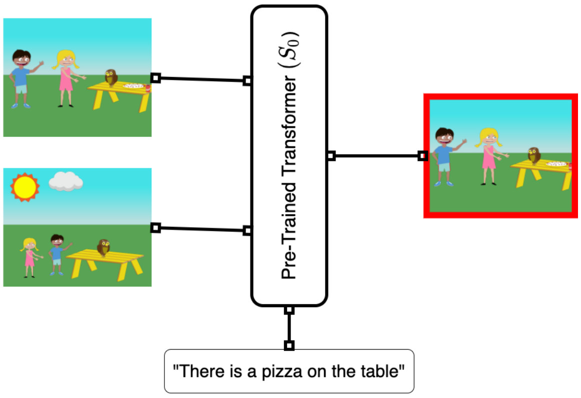
(b) **Rational Listener** L_1: for each pair of images and an input caption, the Rational Listener reuses a pre-trained Transformer Decoder as in S_0 to figure out which image does the caption ground better in. Inspired the teacher-forcing caption training procedure, given an image and the input caption, if it generates a sentence that's closer to the input caption than the other image, then this image is chosen as the target.

(c) **Rational Speaker** S_1: for a pair of images, and a list of candidate captions generated by S_0, the Rational Speaker goes through each candidate caption via the internal simulated listener (same model as L_1 with no disparity), to figure out whether the caption can help the listener pick the correct target image, and if so, how confident. It ranks all the captions by the correctness and confidence score.

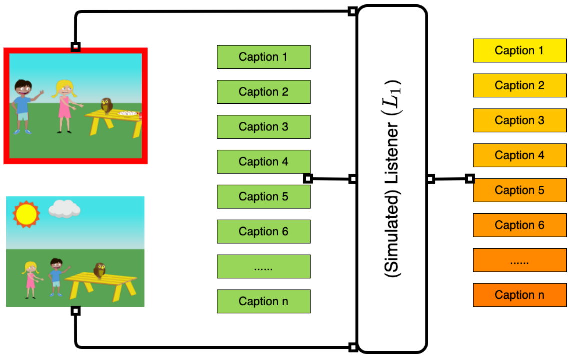
(d) **Pragmatic Rational Speaker** S_1^d: given a list of ranked (by S_1) candidate captions, the Pragmatic Rational Speaker picks the most confident one and send it to the actual listener with disparities (L_1^d), and receives a reward feedback. This feedback helps S_1^d to learn the disparity, and rerank all the captions to accommodate the difference and optimize for the task goal.



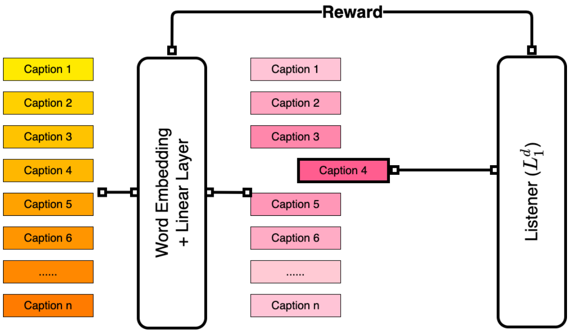
(a) Literal Speaker S_0



(b) Rational Listener L_1/L_1^d



(c) Rational Speaker S_1



(d) Pragmatic Rational Speaker S_1^d

Figure 8: Speaker and Listener model breakdown

Hypernym	Object	Hypernym	Object
boy	mike_reach	girl	jenny_reach
	mike_kick		jenny_kick
	mike_run		jenny_run
	mike_sit		jenny_sit
	mike_fall over		jenny_fall over
	mike_wave		jenny_wave
	mike_up		jenny_up
clothing	blue hat	large objects	bee
	crown		slide
	chef hat		sand
	pirate hat		grill
	sweater hat		swing
	silly hat		tent
	wizzard hat		bench
	horn hat		christmas tree
	glasses		tree
	sunglasses		apple tree
	baseball	food	pie
	glove		pizza
	shovel		hotdog
	racket		ketchup
	kite		mustard
	fire		burger
	bucket		coke
toys	colorful ball	sky objects	helicopter
	basketball		hotair balloon
	soccer		cloud
	tennis ball		sun
	football		lightening
	frisbee		rain
	baseball poll		rocket
	balloon	.	plane
animal	bear	animal	duck
	cat		owl
	dog		snake

Table 3: List of objects and their hypernyms

B Implementation Details

We pretrained the image captioning models using 2 layers of Transformer Decoder with 4 attention heads each, and 512 in internal dimension for 100 epochs each. The dropout rate was 0.5, learning rate started at $1e^{-4}$, on a scheduled decline rate of 0.8 for each 20 unimproved epochs.

We also pretrained the literal listeners and the literal speaker with different disparity datasets. All the weights are fixed before being integrated into the interactive learning phase. During disparity learning, each pair of speaker and listener were trained for 150 epochs, with batch size of 128, learning rate starting at $1e^{-3}$, and on decline at the rate of 0.8 per 20 unimproved epoches. Each experiment is repeated 3 times. The mean and standard deviation were reported in figures. Similarly in the word level training, the model was trained for 200 epochs, and with learning rate starting at 2, and on a scheduled decline rate of 0.8 for each 50 unimproved epochs.

For the efficiency comparison experiment, we used the *combined* test dataset for this experiment, trained each component until 50 unimproved epochs, and selected the top performances within the first 30 minutes of each. All models have reached stable performance by then. All experiments done on a single NVIDIA(R) GeForce(R) RTX 2070 SUPER(TM) 8GB GDDR6 and 10th Gen Intel(R) Core(TM) i9-10900K processor.