A. Formal Version of Theorem 1

Theorem 2. *In one epoch of Proc. 1, if the ToM model is ϵ-optimal, i.e.*

$$\mathcal{L}^{pred} = \mathbb{E}_{s,m} KL[\mathcal{P}_{ToM}(a \mid m, s; \theta) \| P_{l_i}(a \mid o, m)] < \epsilon$$

where states $s = \langle i, k, \mathcal{D}_{supp}, o, m, g \rangle$ and instructions m are sampled as Proc. 1, and for almost all states s speaker gives a δ-optimal instruction candidates pool M, i.e.

$$\sum_{m \in M} \mathcal{P}_{ToM}(a^g \mid m, s; \theta) \geq \delta$$

then expected KL-divergence

$$\mathbb{E}_s KL[\mathcal{Q}_{ToM}(m \mid s) \| \mathcal{Q}(m \mid s; \theta)] \tag{15}$$

between the instruction distribution calculated from ToM model

$$\mathcal{Q}_{ToM}(m \mid s; \theta) \triangleq \frac{\mathcal{P}_{ToM}(a^g \mid m, s; \theta)}{\sum_{m' \in M} \mathcal{P}_{ToM}(a^g \mid m', s; \theta)} \tag{16}$$

and the target instruction distribution

$$\mathcal{Q}(m \mid s) \triangleq \frac{P_{l_i}(a^g \mid o, m)}{\sum_{m' \in M} P_{l_i}(a^g \mid o, m')} \tag{17}$$

upper-bounded by

$$\frac{N_M \sqrt{\frac{\epsilon}{2(1-\delta)}} + W_0(\epsilon)}{\delta} \tag{18}$$

where N_M is the size of largest pool of instruction candidates produced by the speaker, and W_0 is the principle branch of Lambert's W function.

Proof. Applying Pinsker inequality,

$$\begin{aligned}
\mathcal{L}^{\text{pred}} &= \mathbb{E}_{s,m} KL[\mathcal{P}_{\text{ToM}}(a \mid m, s; \theta) \| P_{l_i}(a \mid o, m)] \\
&\geq \mathbb{E}_{s,m} 2TV(P_{l_i}(a \mid o, m), \mathcal{P}_{\text{ToM}}(a \mid m, s; \theta))^2 \\
&= \mathbb{E}_{s,m} 2 \sup_a |P_{l_i}(a \mid o, m) - \mathcal{P}_{\text{ToM}}(a \mid m, s; \theta)|^2 \\
&\geq \mathbb{E}_{s,m} 2 |P_{l_i}(a^g \mid o, m) - \mathcal{P}_{\text{ToM}}(a^g \mid m, s; \theta)|^2 \\
&\geq \mathbb{E}_{s,m} 2 |\Delta(s, m))|^2 \\
&\geq 2(1-\sigma) \mathbb{E}_s \mathbb{E}_{m \sim \mathcal{U}(M)} |\Delta(s, m))|^2 \\
&\geq 2(1-\sigma) (\mathbb{E}_s \mathbb{E}_{m \sim \mathcal{U}(M)} |\Delta(s, m))|)^2
\end{aligned} \tag{19}$$

where $\Delta(s, m) = P_{l_i}(a^g \mid o, m) - \mathcal{P}_{\text{ToM}}(a^g \mid m, s; \theta)$.

Model	Ave success (%)
Gold-standard speaker	91.20
Non-ToM speaker	37.38
RSA w/ single listener	39.32
RSA speaker	42.83
Finetuned RSA	44.30
ToM. speaker (large $h = 768$)	55.28
ToM. speaker (small $h = 256$)	56.75
ToM. speaker ($N_{\text{inner}} = 1$)	56.10
ToM. speaker ($N_{\text{inner}} = 10$)	58.25
ToM. speaker	**58.19**

Table 2. The influence of various hyperparameters

By processing the target expectation

$$\begin{aligned}
&\mathbb{E}_s KL[\mathcal{Q}_{\text{ToM}}(m \mid s; \theta) \| \mathcal{Q}(m \mid s)] \\
=&\mathbb{E}_s \log \frac{\sum_{m' \in M} P_{l_i}(a^g \mid o, m')}{\sum_{m' \in M} \mathcal{P}_{\text{ToM}}(a^g \mid m', s; \theta)} \\
&+ \mathbb{E}_s \frac{\sum_{m \in M} \log \frac{\mathcal{P}_{\text{ToM}}(a^g \mid m, s; \theta)}{P_{l_i}(a^g \mid o, m)} \mathcal{P}_{\text{ToM}}(a^g \mid m, s; \theta)}{\sum_{m' \in M} \mathcal{P}_{\text{ToM}}(a^g \mid m', s; \theta)} \\
\leq& \frac{N_M}{\delta} \mathbb{E}_s \mathbb{E}_m \Delta(s, m') \\
&+ \mathbb{E}_s \frac{\sum_{m \in M} W_0(KL[\mathcal{P}_{\text{ToM}}(a \mid m, s; \theta) \| P_{l_i}(a \mid o, m)])}{\delta} \\
=& \frac{N_M \sqrt{\frac{\epsilon}{2(1-\delta)}} + W_0(\epsilon)}{\delta}
\end{aligned} \tag{20}$$

□

B. Training Time and space

All of our models can be trained on a 32 Gb V100. A model (speaker, listener, or ToM model) for referential game trains for about 20 hours, while a model (speaker, listener, or ToM model) for language navigation trains for 72 about hours. Tab. 1 and Fig. 3 reports the average of three runs, Fig. 2 reports data from 20 testing listeners.

C. Hyper-parameter Tuning

We only tuned the inner and outer learning rates of MAML among $1e^i, i = -1, -2, -3, -4, -5$. A few influential hyperparameters are shown in Tab. 2. Other parameters are all kept same as previous work: Lowe et al. (2019a) for referential game, and Shridhar et al. (2021) for language navigation.