Interpreting Emergent Communication

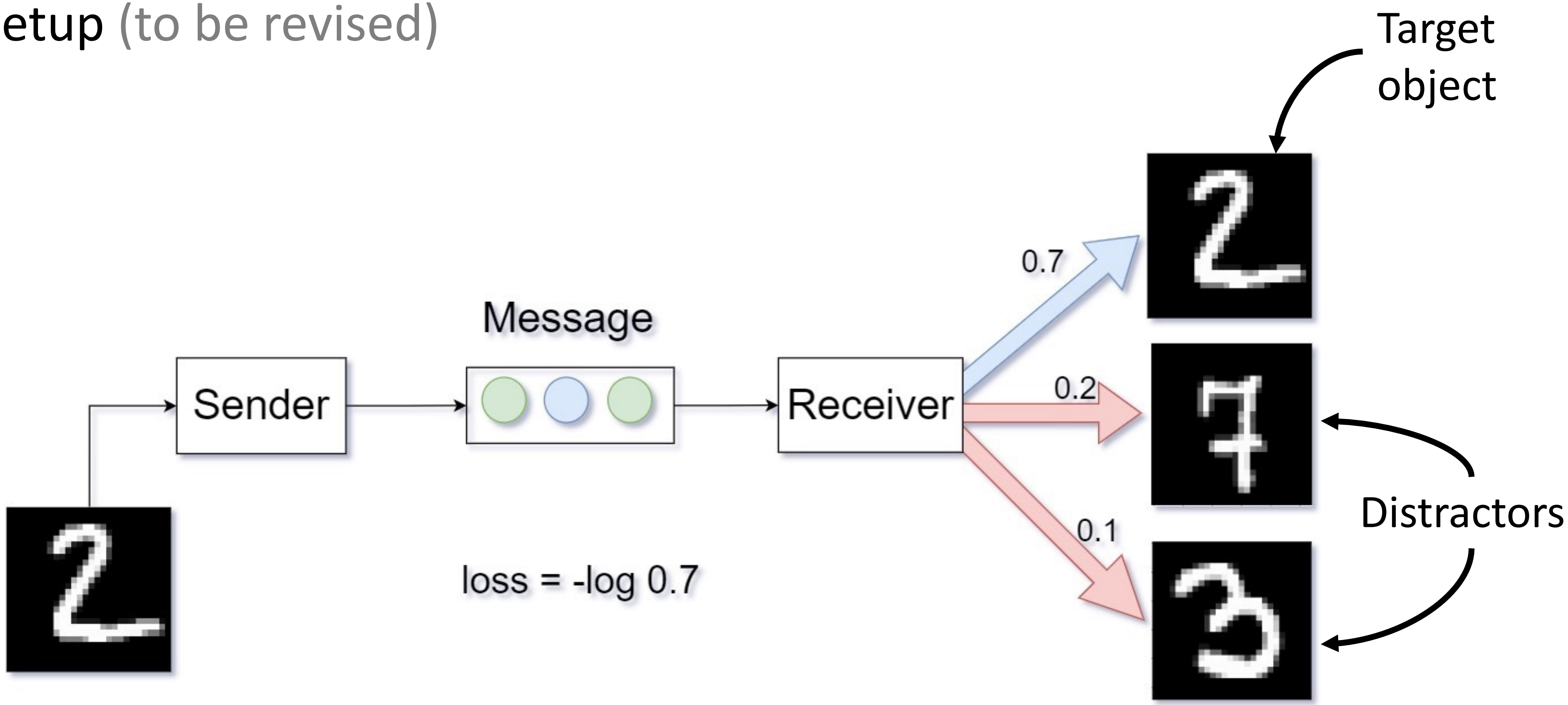
Yonatan Belinkov

Decoding Communication in Non-Human Species III
Simons Institute, June 29, 2024

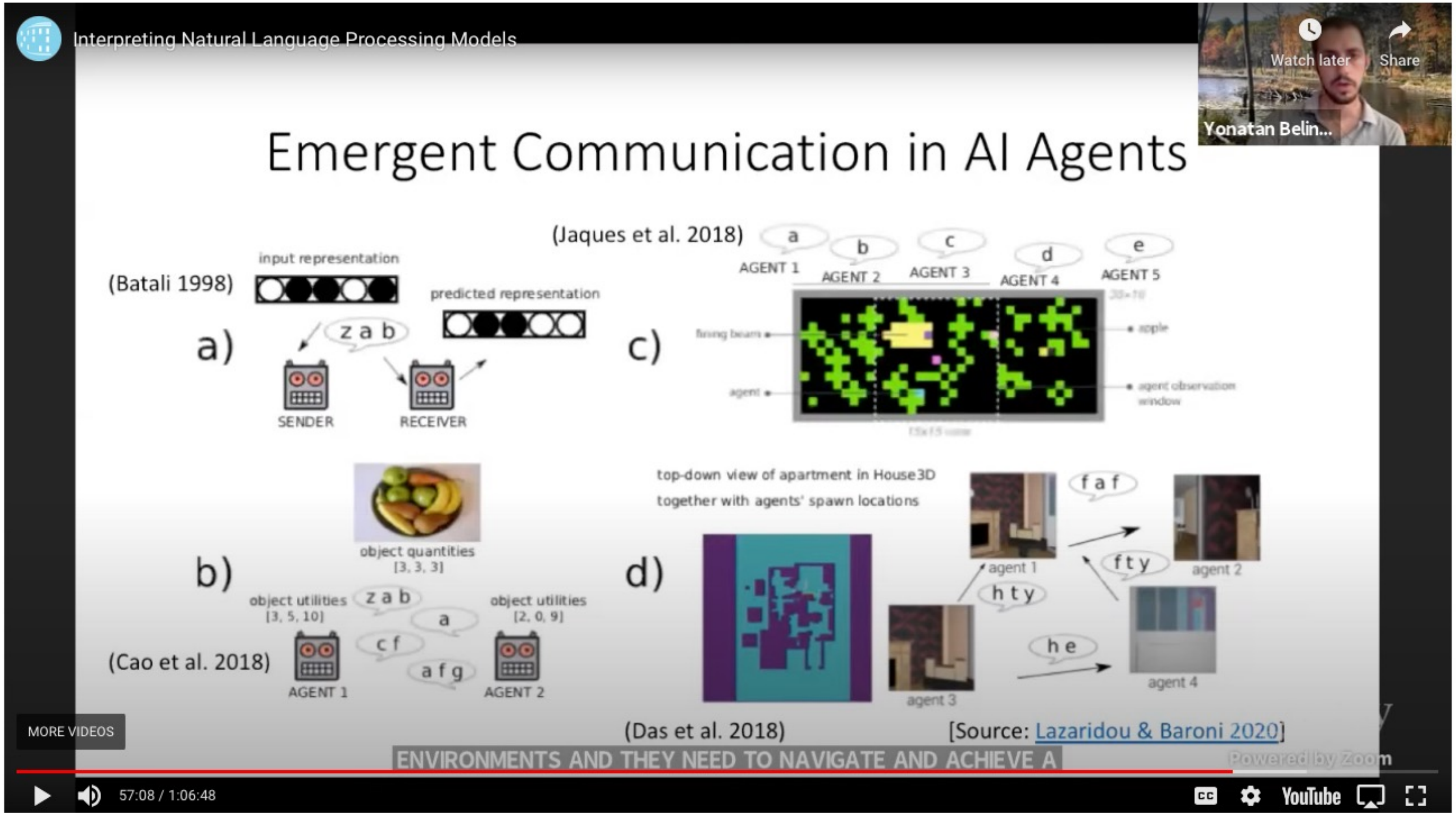
TECHNION | The Henry and Marilyn Taub Faculty of Computer Science

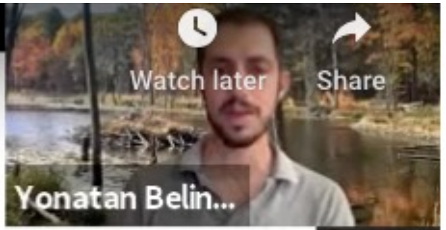
Emergent communication

Basic setup (to be revised)



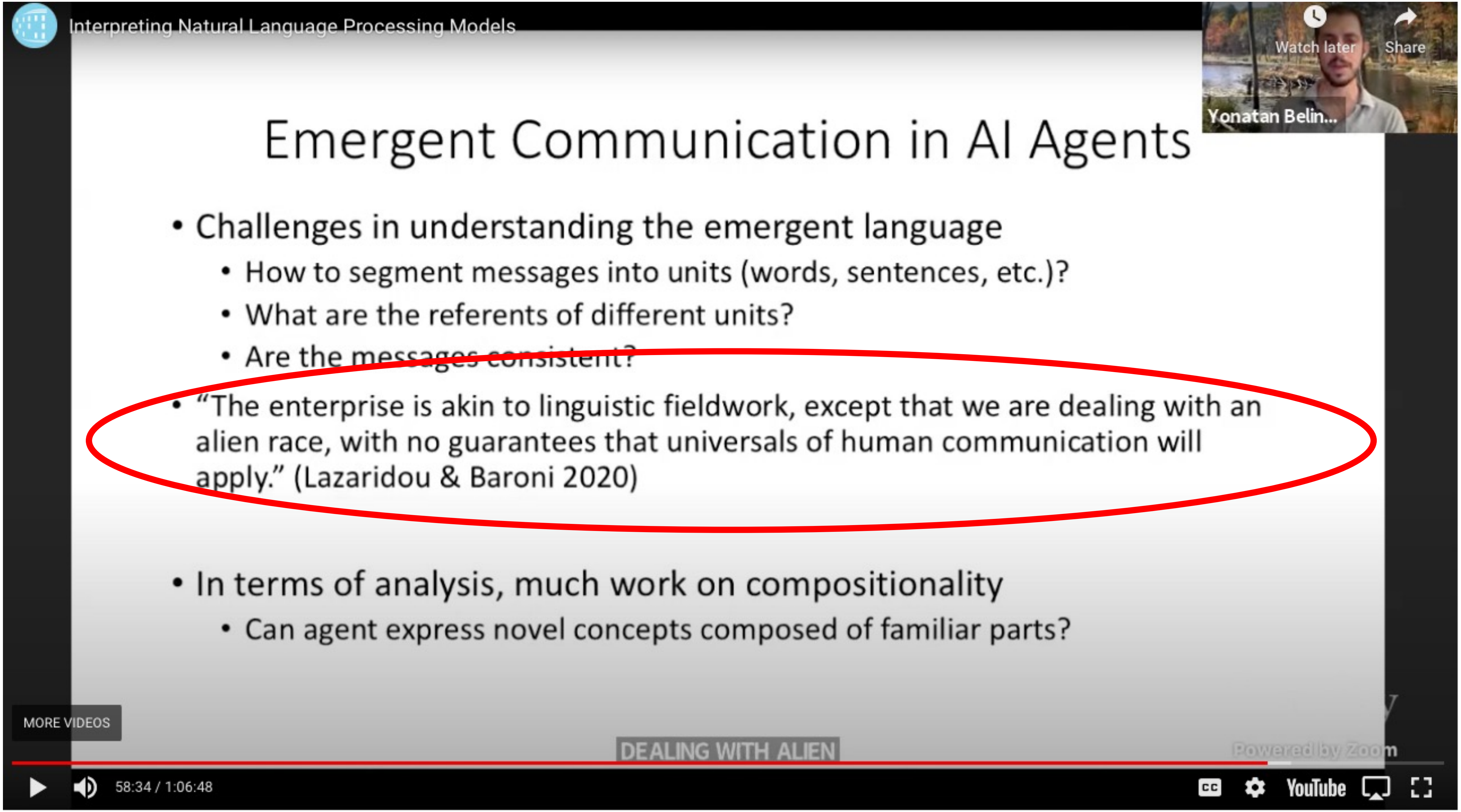
Flashback to 2020: iteration 1 of this workshop





Emergent Communication in AI Agents

- Challenges in understanding the emergent language
  - How to segment messages into units (words, sentences, etc.)?
  - What are the referents of different units?
  - Are the messages consistent?
- "The enterprise is akin to linguistic fieldwork, except that we are dealing with an alien race, with no guarantees that universals of human communication will apply." (Lazaridou & Baroni 2020)
- In terms of analysis, much work on compositionality
  - Can agent express novel concepts composed of familiar parts?


Interpreting Natural Language Processing Models
Watch later
Share
Yonatan Belin...
Emergent Communication in AI Agents
• Challenges in understanding the emergent language
• How to segment messages into units (words, sentences, etc.)?
• What are the referents of different units?
• Are the messages consistent?
• "The enterprise is akin to linguistic fieldwork, except that we are dealing with an alien race, with no guarantees that universals of human communication will apply." (Lazaridou & Baroni 2020)
• In terms of analysis, much work on compositionality
• Can agent express novel concepts composed of familiar parts?
MORE VIDEOS
DEALING WITH ALIEN
Powered by Zoom
58:34 / 1:06:48
YouTube

How to interpret the emergent communication?

Most common: just evaluate accuracy on the task

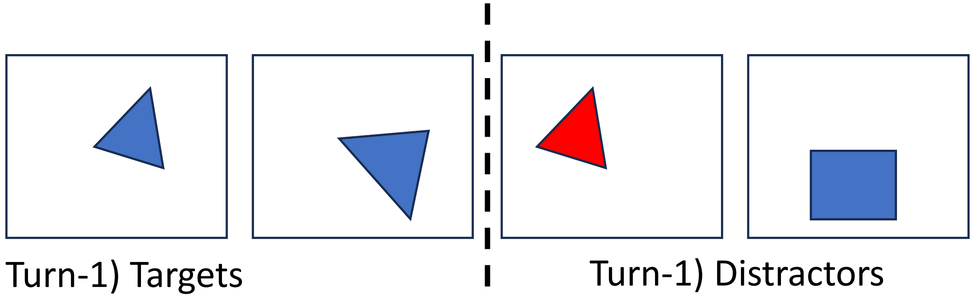
Compositionality of emergent language?

- Topographic similarity (global, opaque, not correlated with accuracy)

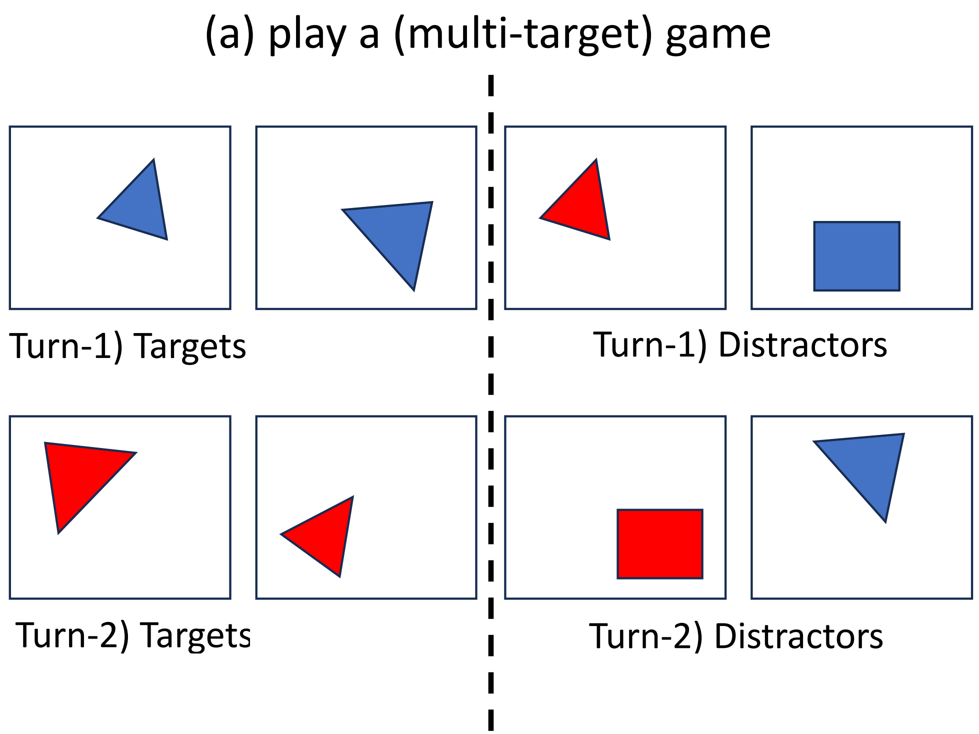
We want: interpretable, specific metric

How to interpret the emergent communication?

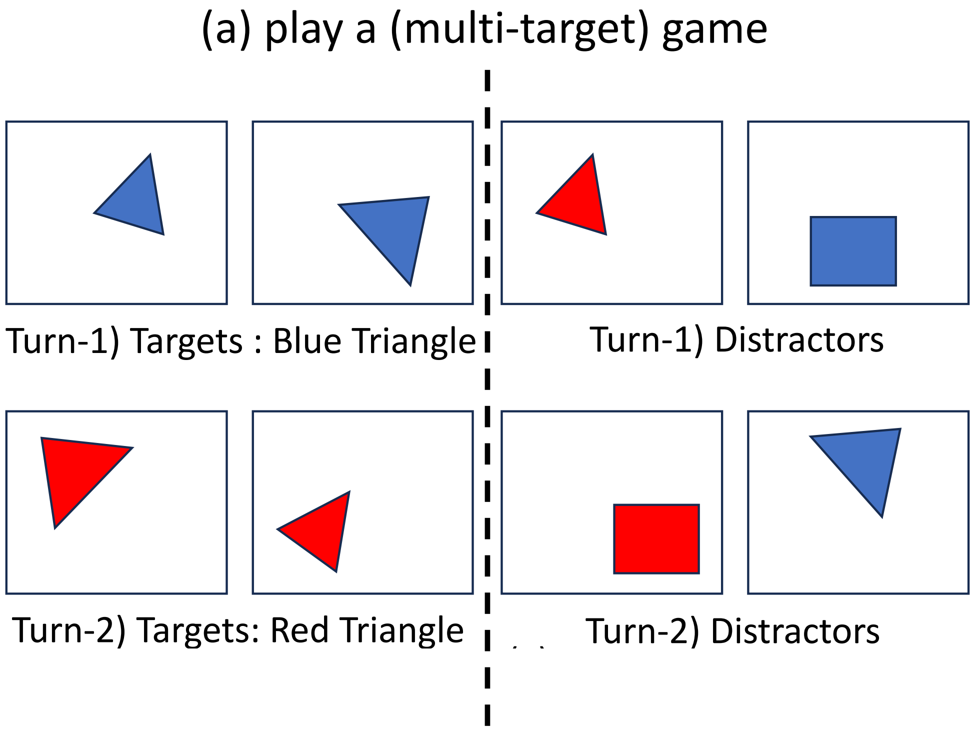
(a) play a (multi-target) game



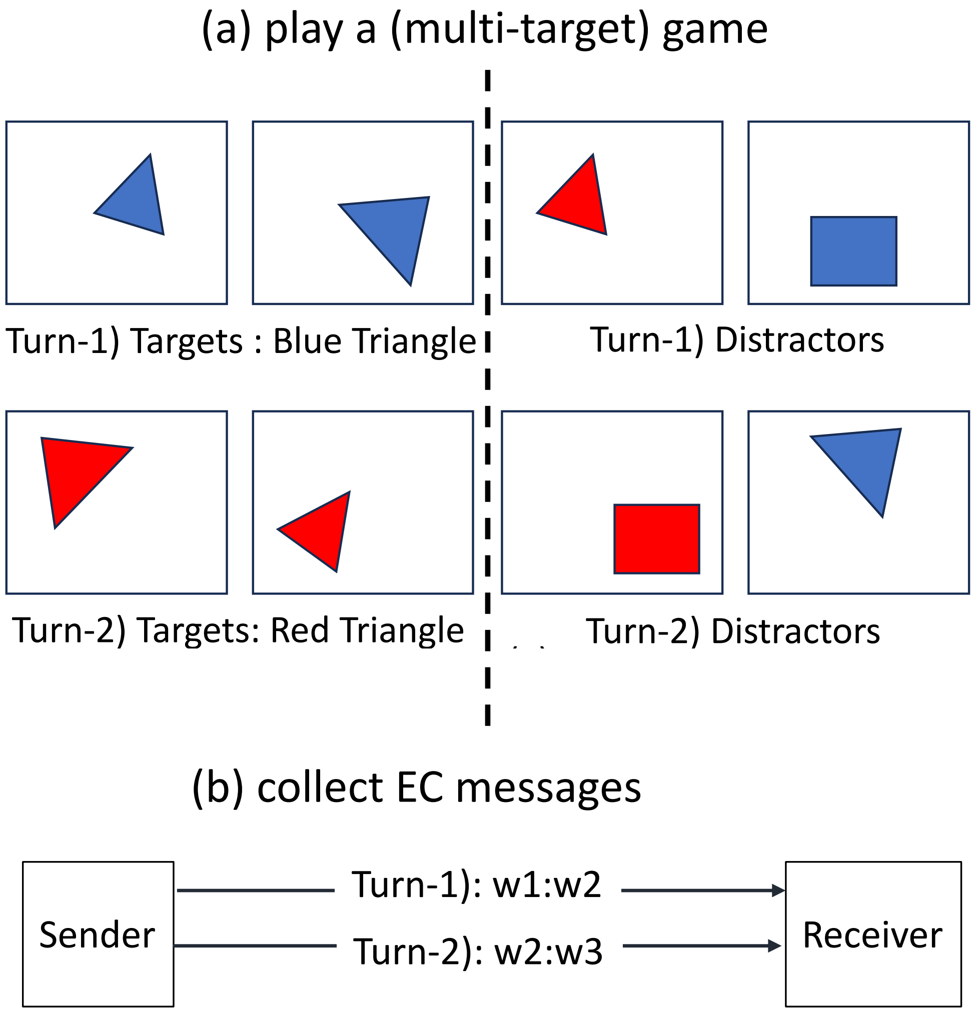
How to interpret the emergent communication?



How to interpret the emergent communication?

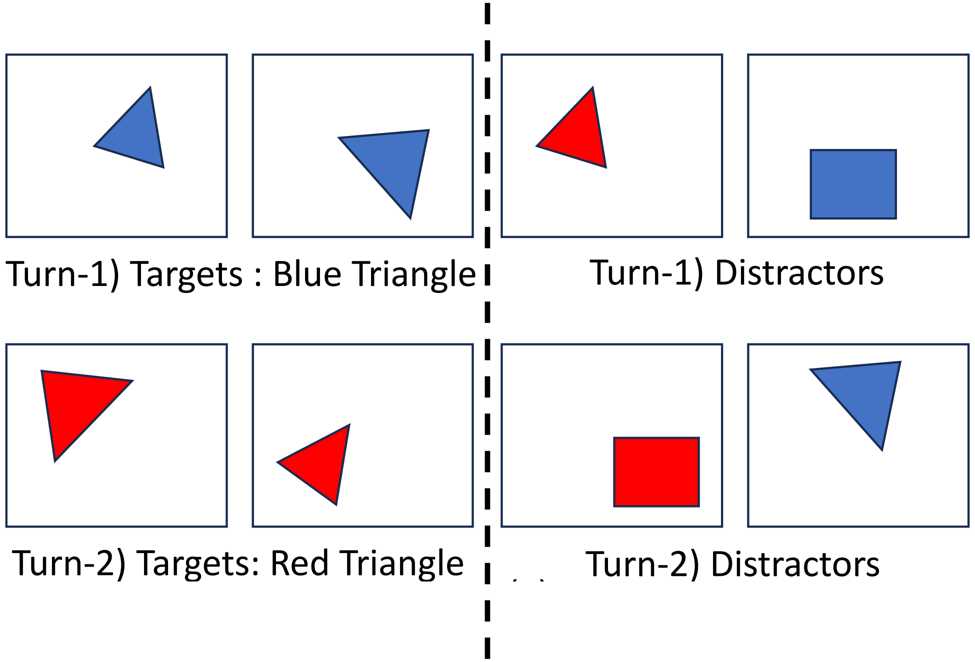


How to interpret the emergent communication?

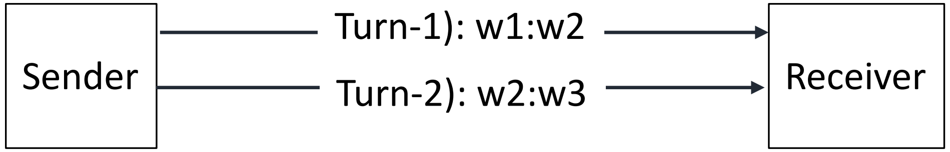


How to interpret the emergent communication?

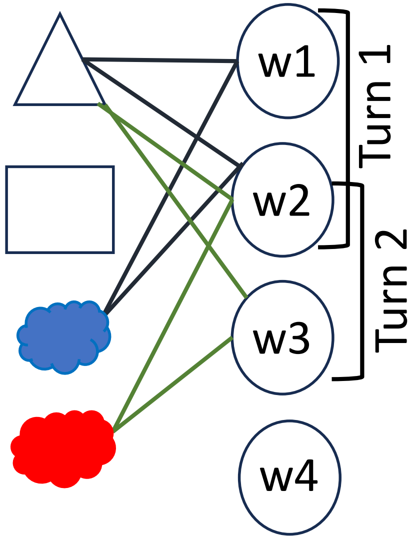
(a) play a (multi-target) game



(b) collect EC messages

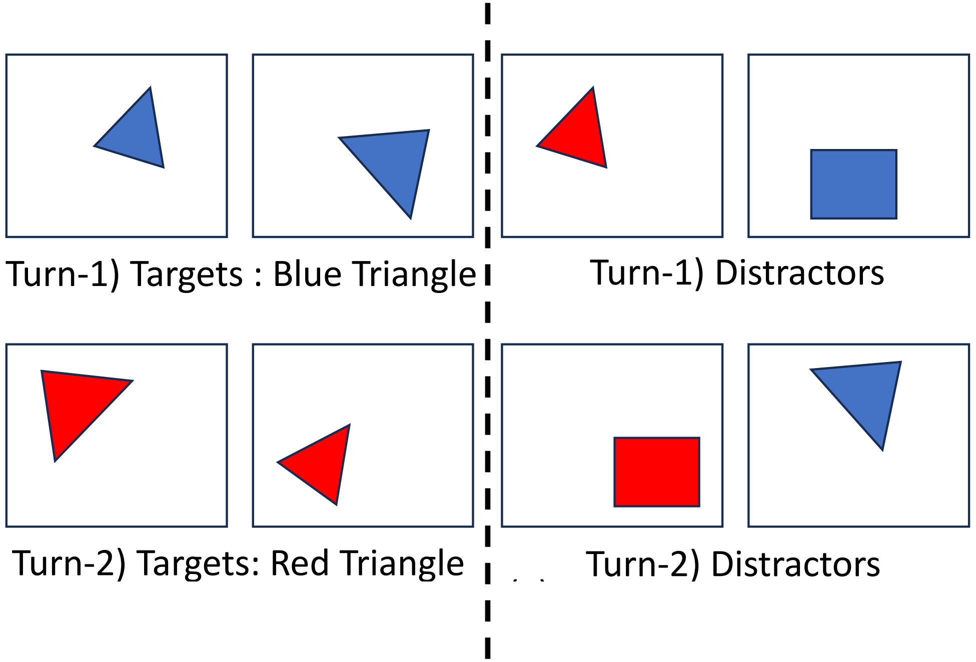


(a) bi-partite graph of concepts and messages

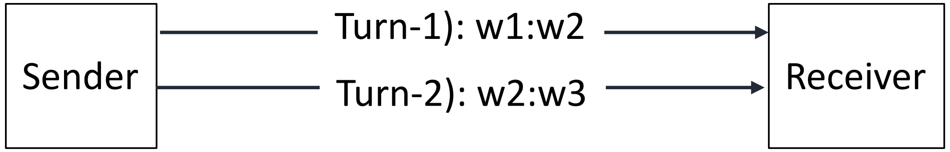


How to interpret the emergent communication?

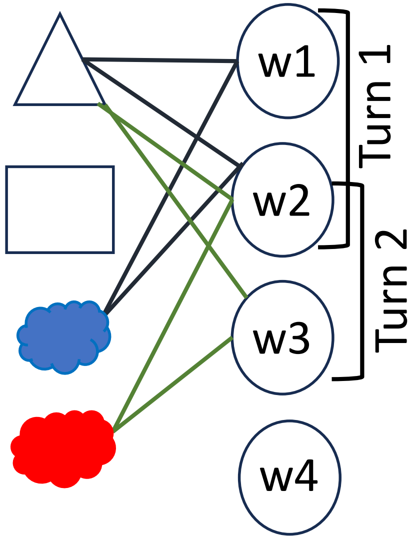
(a) play a (multi-target) game



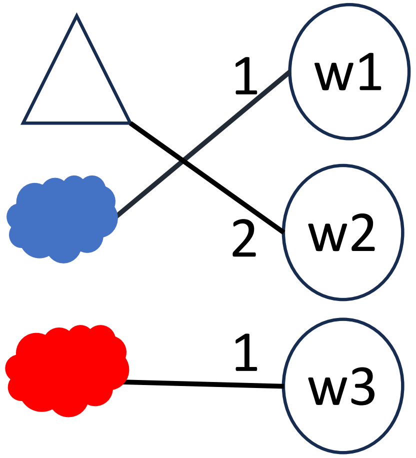
(b) collect EC messages



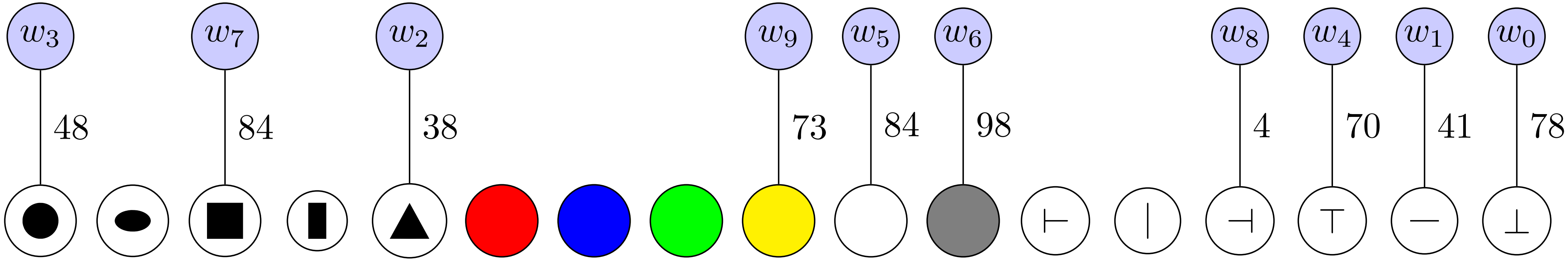
(a) bi-partite graph of concepts and messages



(d) best match



Best-match graph



Desiderata from a communication system

Works well – agents succeed in the task

Interpretable – humans can understand it (“good best-match graph”)

Supports compositional generalization

Desiderata from a communication system

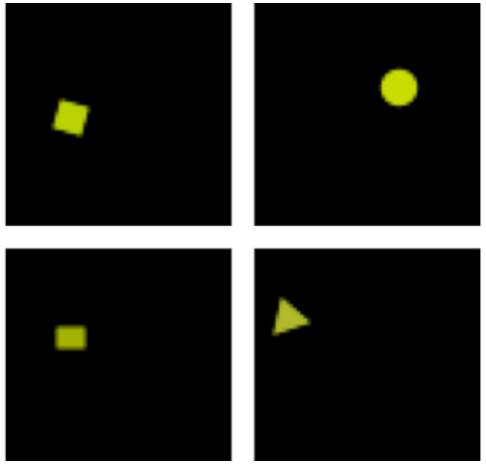
Works well – agents succeed in the task

Interpretable – humans can understand it

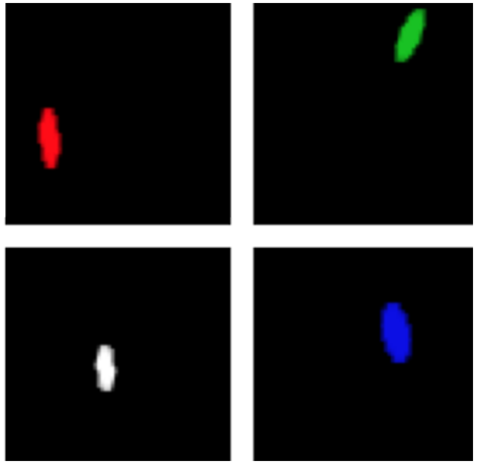
Supports compositional generalization

Train on **single** concepts

“yellow”

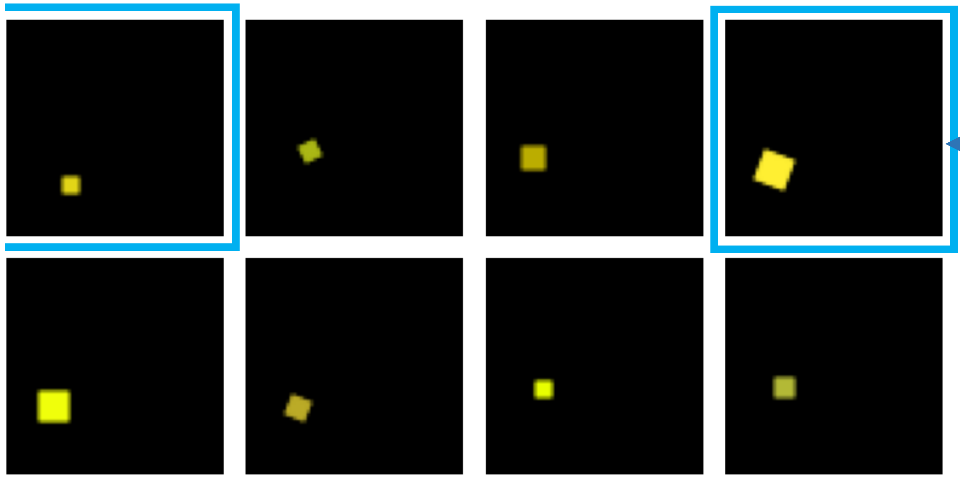


“ellipse”



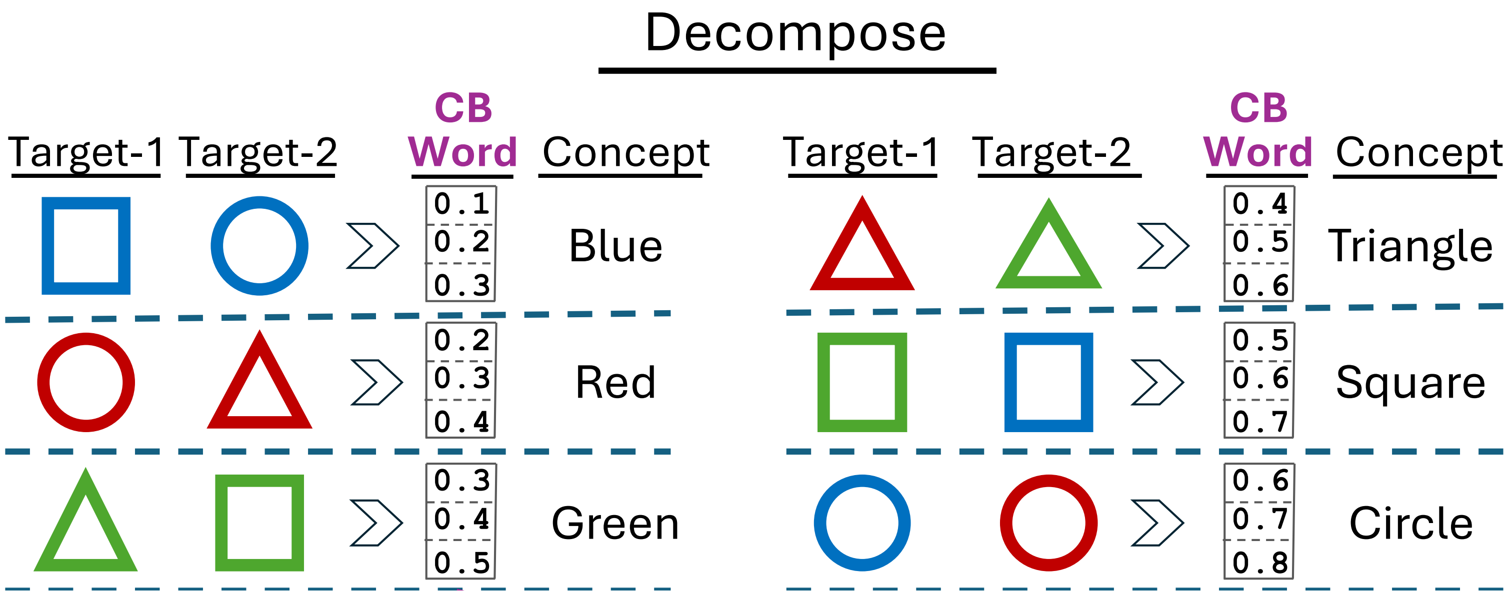
Succeed in **composite** phrases

“yellow square bottom left”



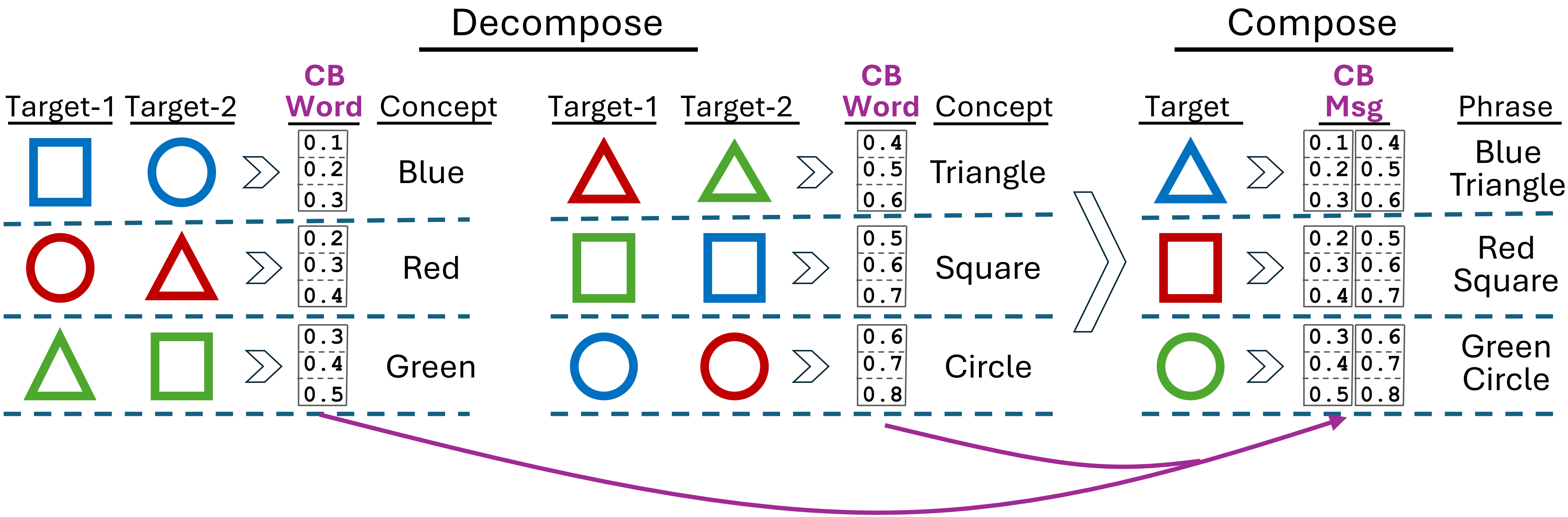
Composition through Decomposition

“Break down to build up”

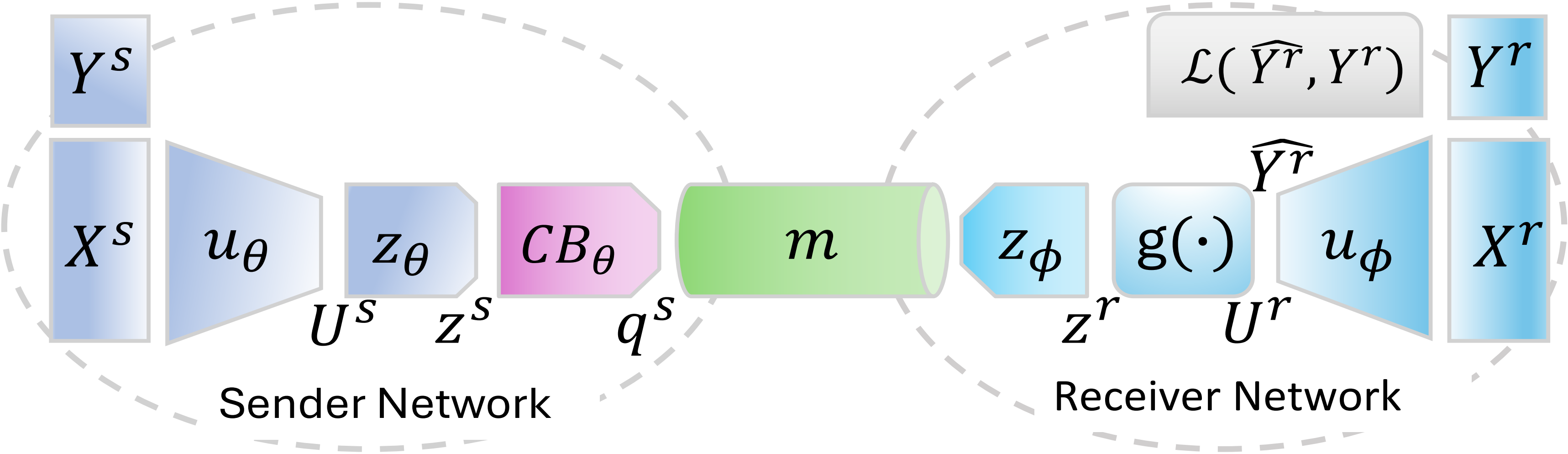


Composition through Decomposition

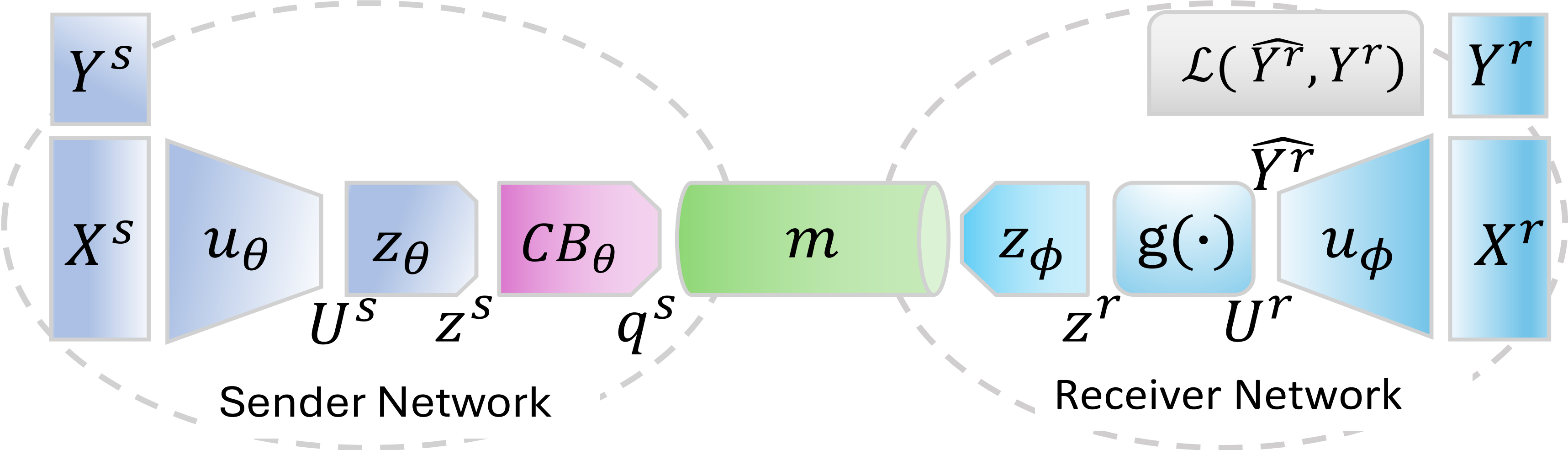
"Break down to build up"

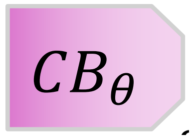


Communication with a discrete codebook



Communication with a discrete codebook



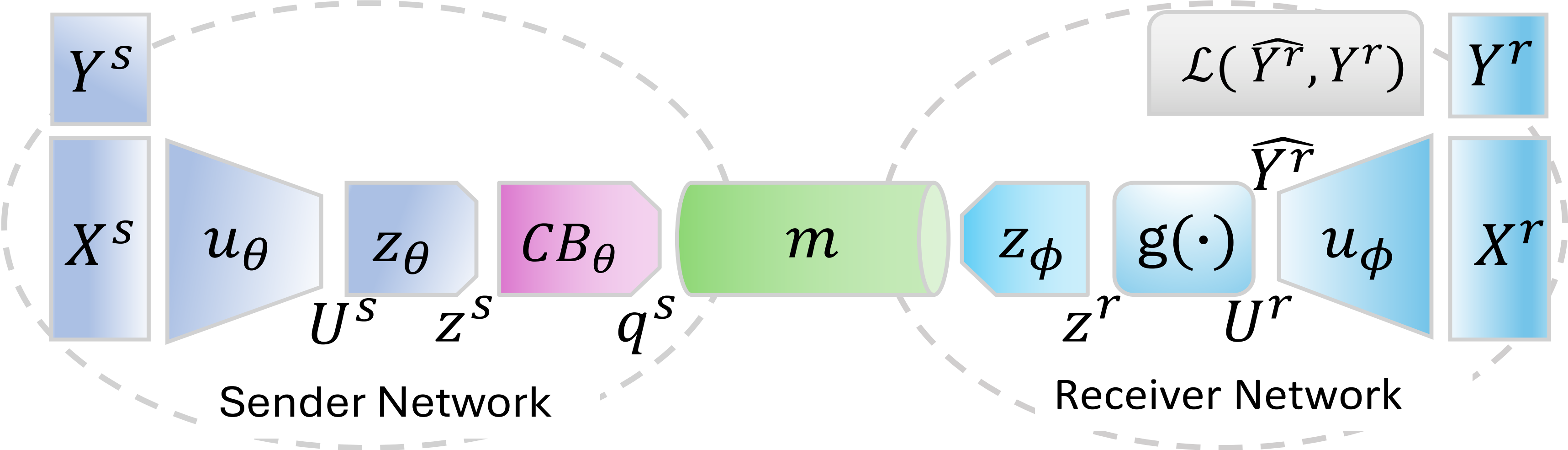


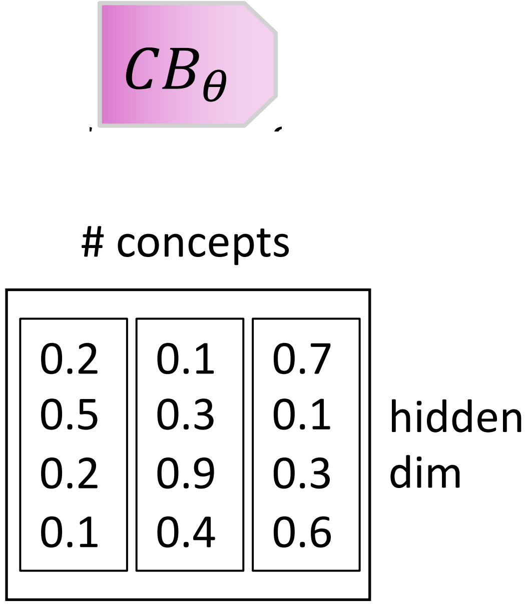
concepts

0.2	0.1	0.7
0.5	0.3	0.1
0.2	0.9	0.3
0.1	0.4	0.6

hidden dim

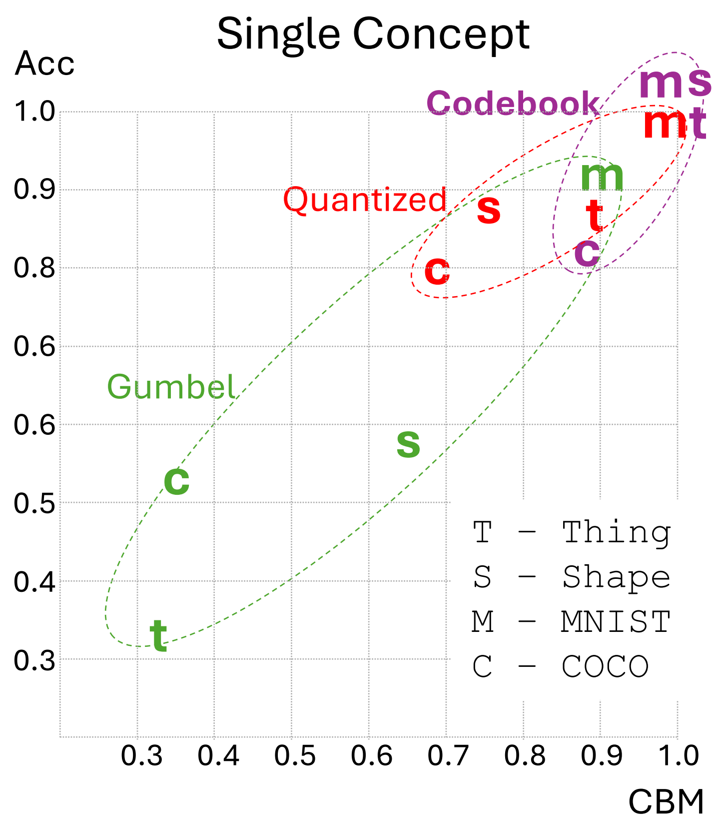
Communication with a discrete codebook

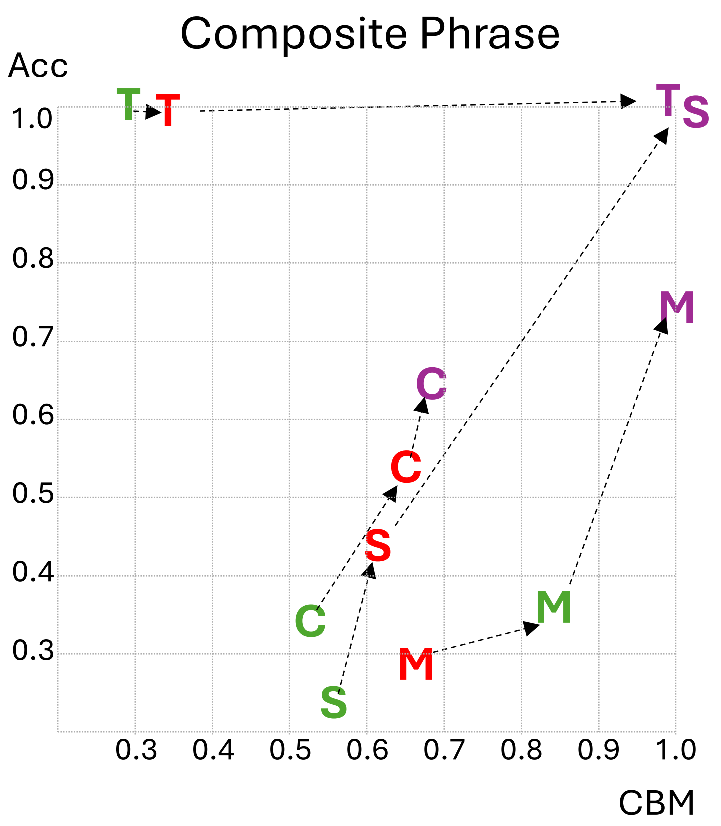




- Codebook training based on ideas from VQ-VAEs
- Objective balances task performance with codebook 'quality'

Composition through Decomposition works





Zero-shot compositional generalization

	Thing		Shape		Mnist		Coco		Qrc	
Method	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM
C/D	0.25	0.25	0.26	0.47	0.41	0.58	0.61	0.56	0.98	0.17
CtD	1.00	1.00	0.99	1.00	0.81	0.95	0.66	0.64	0.95	0.17
CtD ZS	1.00	1.00	0.81	1.00	0.89	0.96	0.74	0.68	0.48	0.52

Zero-shot: no training on composite phrases

Zero-shot compositional generalization

	Thing		Shape		Mnist		Coco		Qrc	
Method	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM
C/D	0.25	0.25	0.26	0.47	0.41	0.58	0.61	0.56	0.98	0.17
CtD	1.00	1.00	0.99	1.00	0.81	0.95	0.66	0.64	0.95	0.17
CtD ZS	1.00	1.00	0.81	1.00	0.89	0.96	0.74	0.68	0.48	0.52

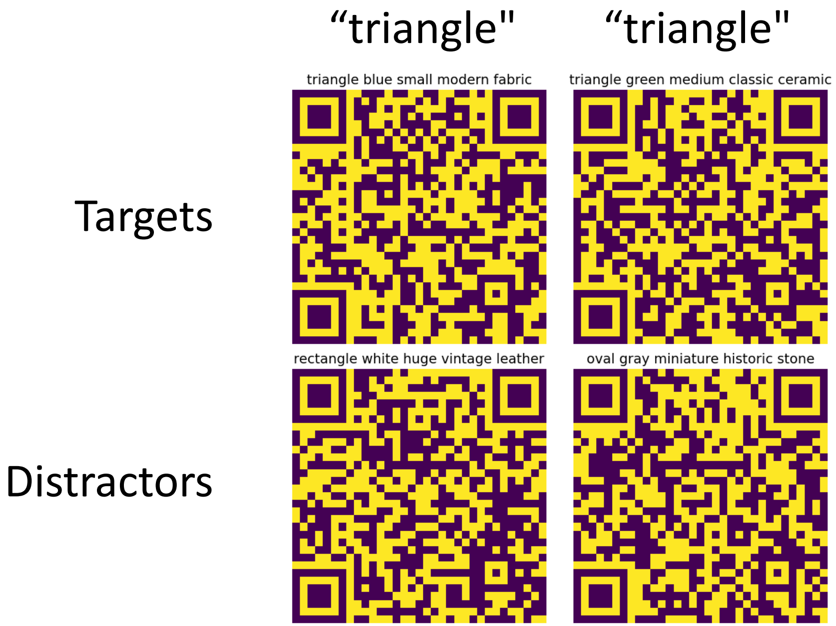
Zero-shot: no training on composite phrases

Zero-shot compositional generalization

	Thing		Shape		Mnist		Coco		Qrc	
Method	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM	Acc	CBM
C/D	0.25	0.25	0.26	0.47	0.41	0.58	0.61	0.56	0.98	0.17
CtD	1.00	1.00	0.99	1.00	0.81	0.95	0.66	0.64	0.95	0.17
CtD ZS	1.00	1.00	0.81	1.00	0.89	0.96	0.74	0.68	0.48	0.52

Zero-shot: no training on composite phrases

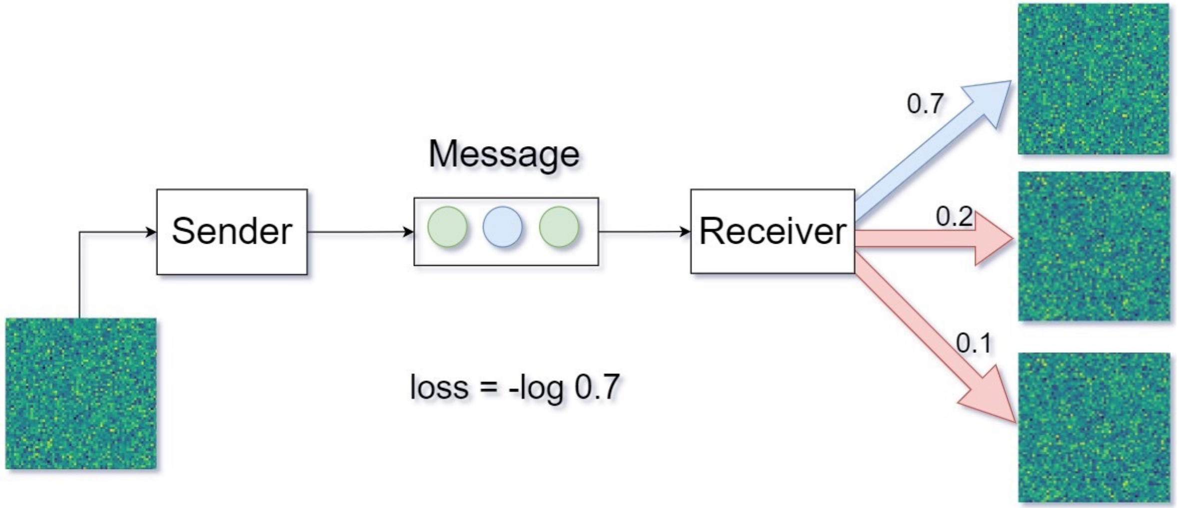
But: when the dataset isn't compositional, we cannot decompose!



What is required for meaningful communication to emerge?

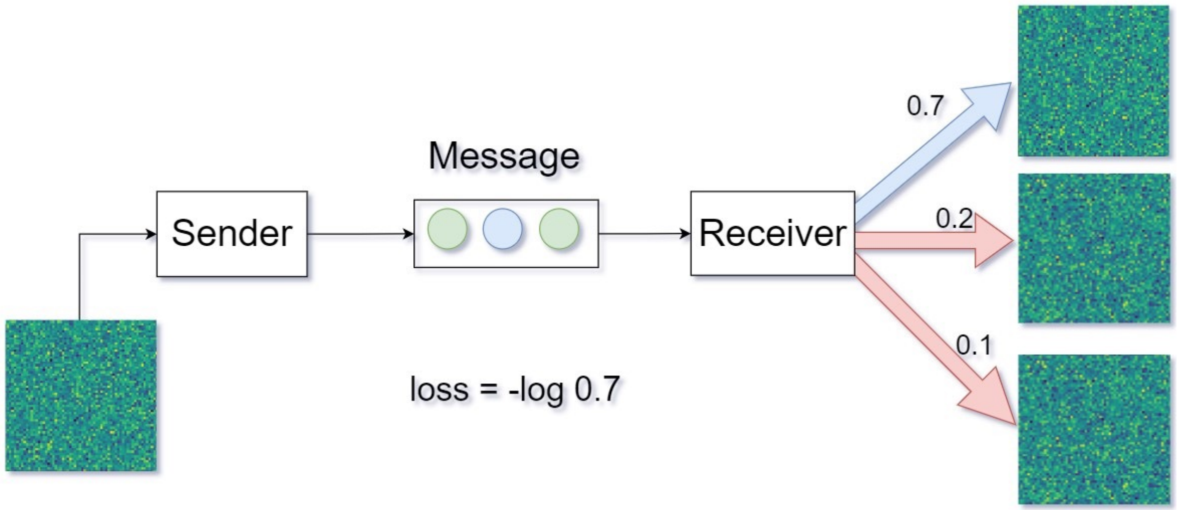
A surprising result

Some EC agents can generalize to noise! (Bouchacourt & Baroni 2018)



A surprising result

Some EC agents can generalize to noise! (Bouchacourt & Baroni 2018)

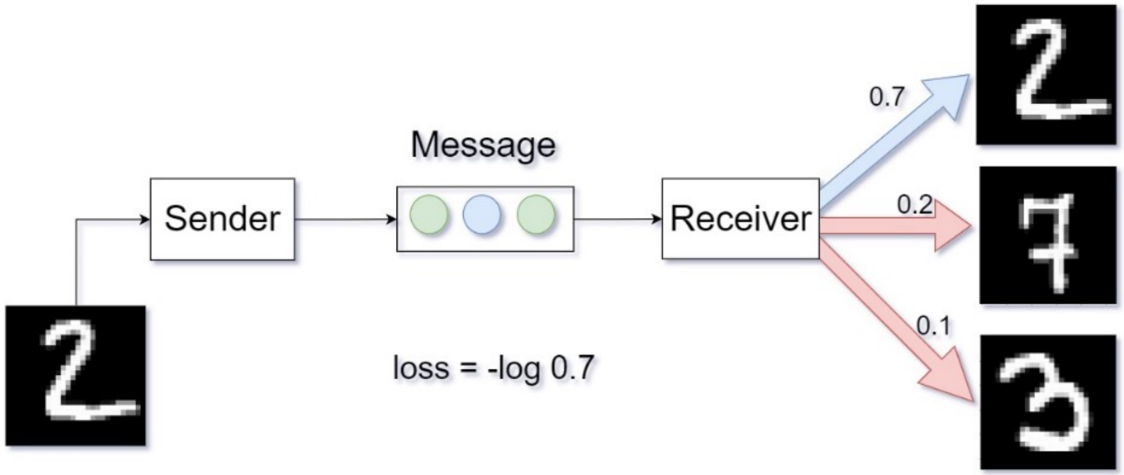


Implication: some properties of natural language are **not necessary** for task success

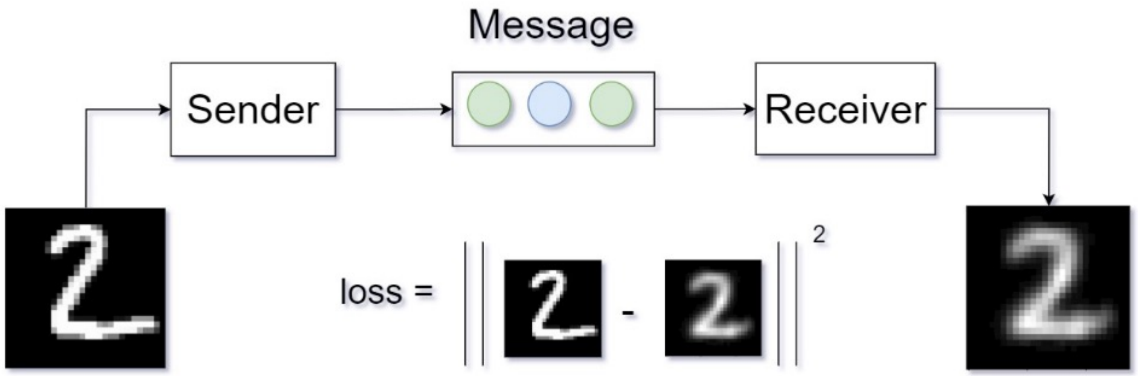
→ What drives **meaningful** communication?

Common EC setups

Discrimination



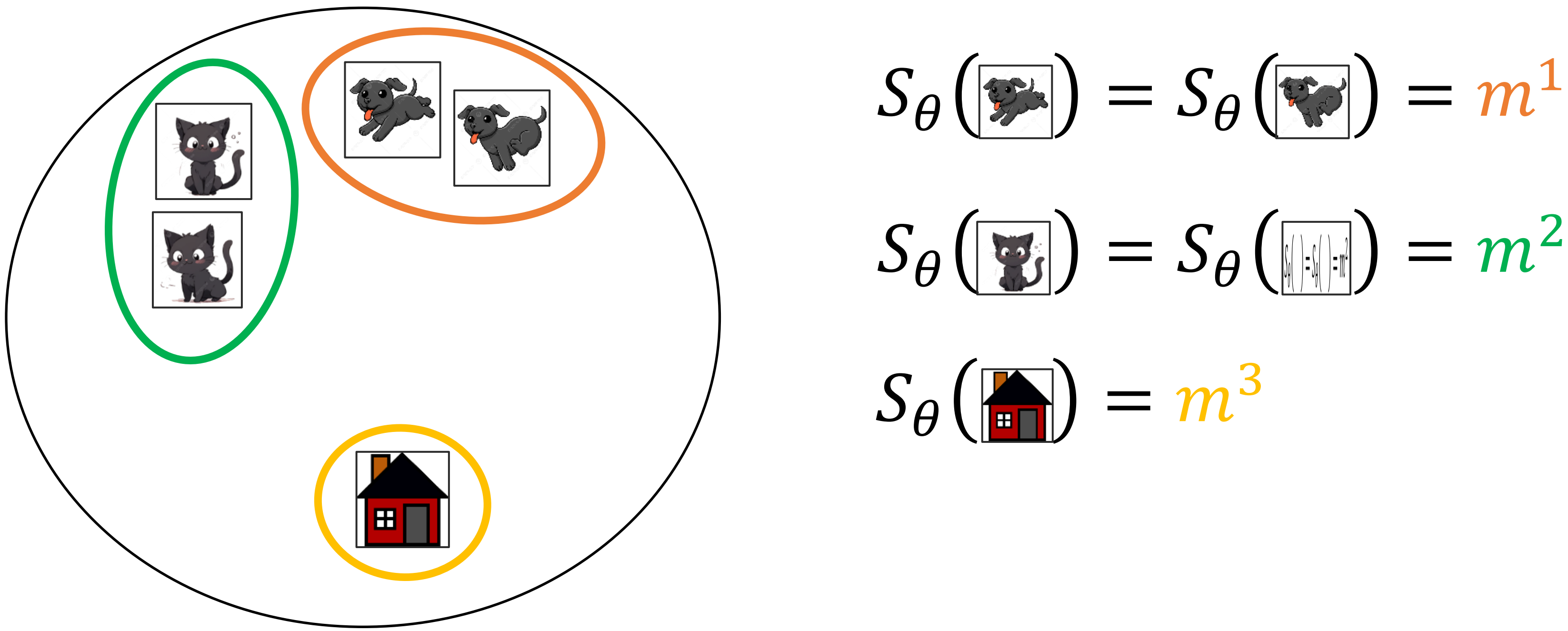
Reconstruction



The communication protocol is a **many-to-one** mapping

Semantic consistency

Similar objects are mapped to the same message



Definition: Semantic consistency

A communication protocol S_θ is **semantically consistent** if

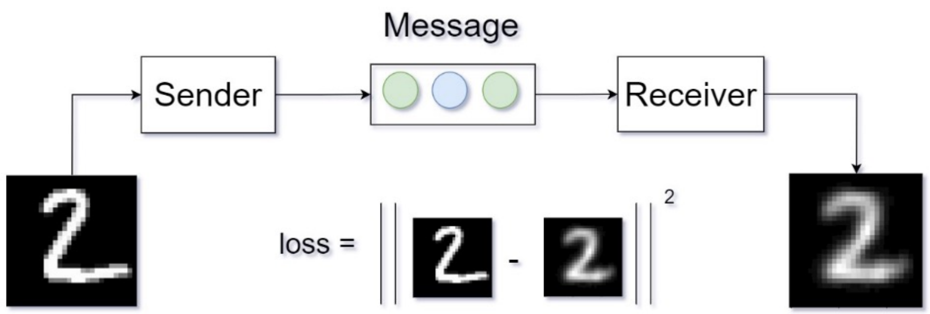
$$\mathop{\mathbb{E}}_{m \sim S_\theta(X)} \left[\mathrm{Var}\left[X \mid S_\theta(X) = m\right]\right] < \mathrm{Var}\left[X\right]$$

Equivalently:

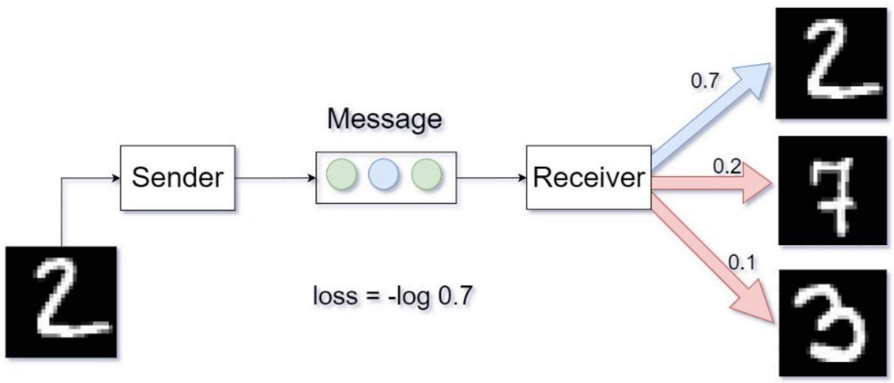
$$\mathop{\mathbb{E}}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2 \;\middle|\; S_\theta(x_1) = S_\theta(x_2)\right] < \mathop{\mathbb{E}}_{x_1, x_2 \sim X} \left[\|x_1 - x_2\|^2\right]$$

Inputs mapped to the same message are more similar than random inputs

Reconstruction

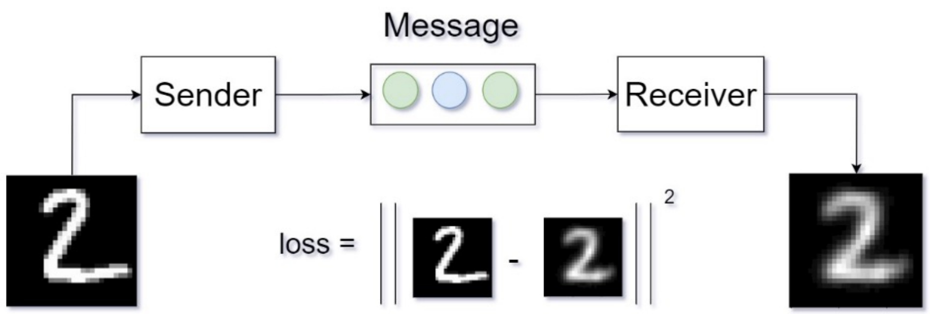


Discrimination



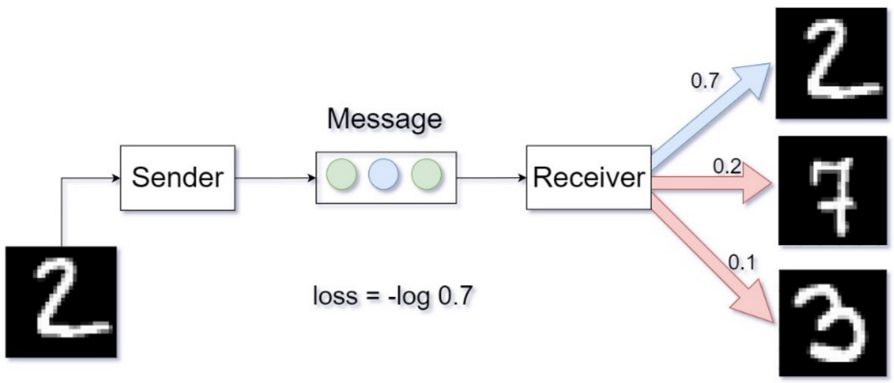
Theorem: Assuming receiver Φ is unrestricted and sender space Θ contains at least one semantically consistent protocol, every optimal communication protocol is semantically consistent

Reconstruction



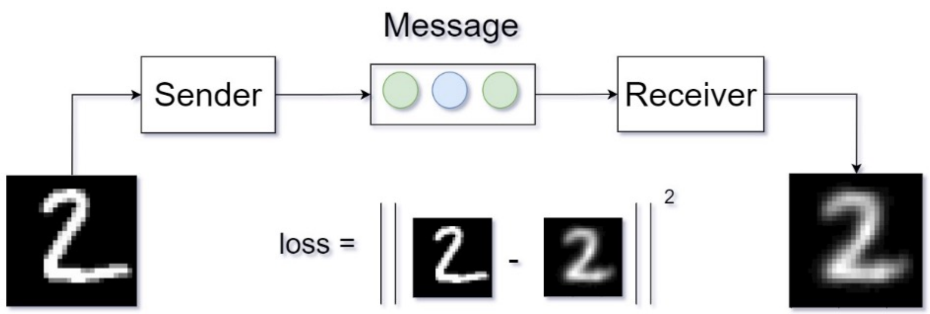
Theorem: Assuming receiver Φ is unrestricted and sender space Θ contains at least one semantically consistent protocol, every optimal communication protocol is semantically consistent

Discrimination



Theorem: There exists a game where receiver Φ is unrestricted and sender space Θ contains at least one semantically consistent protocol, in which not all the optimal communication protocols are semantically consistent.

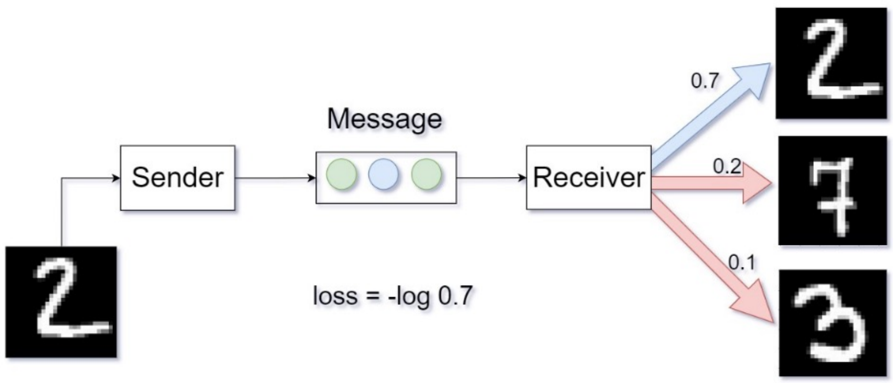
Reconstruction



Theorem: Assuming receiver Φ is unrestricted and sender space Θ contains at least one semantically consistent protocol, every optimal communication protocol is semantically consistent

$$loss^*(S_\theta) = \sum_{m \in M} P(S_\theta(X) = m) \cdot \mathrm{Var}\left[X \mid S_\theta(X) = m\right]$$

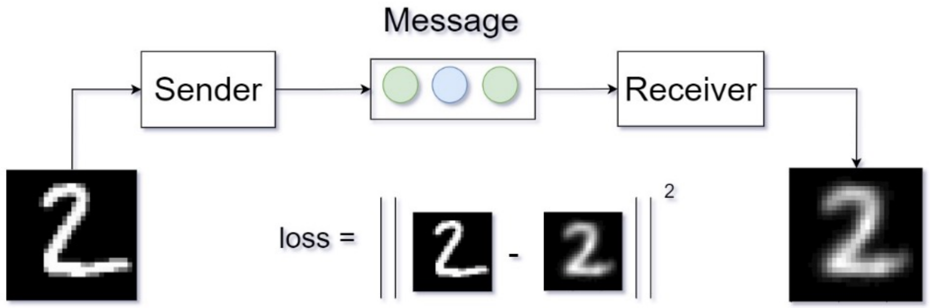
Discrimination

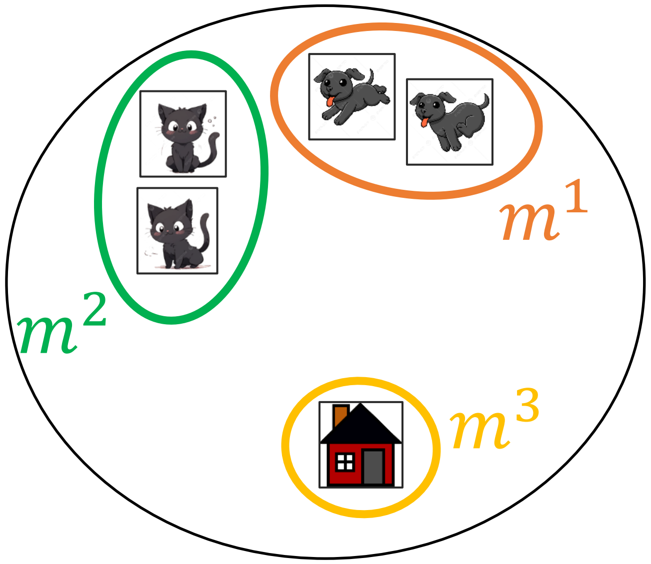


Theorem: There exists a game where receiver Φ is unrestricted and sender space Θ contains at least one semantically consistent protocol, in which not all the optimal communication protocols are semantically consistent.

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

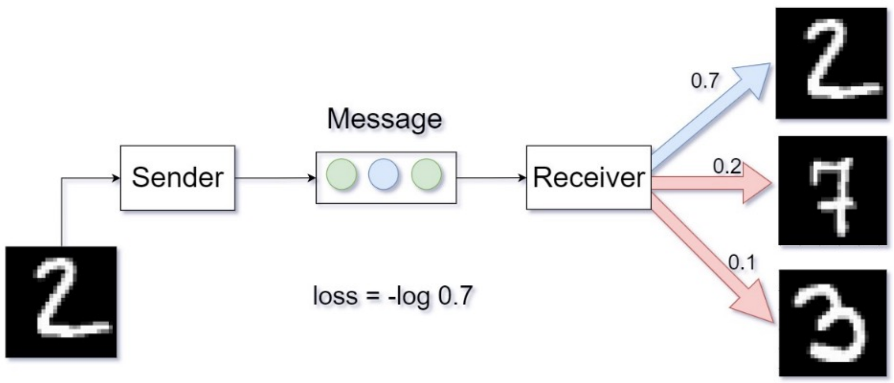
Reconstruction

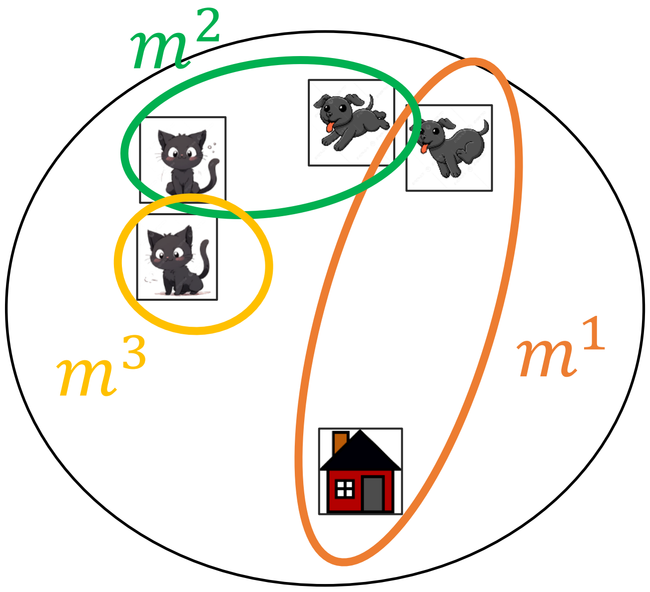




Every optimal communication is consistent

Discrimination





There exists an optimal communication that is not consistent

*Assuming Receiver is perfectly optimized with an unrestricted hypothesis class

Spatial meaningfulness

Semantic consistency requires hard equality of messages

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

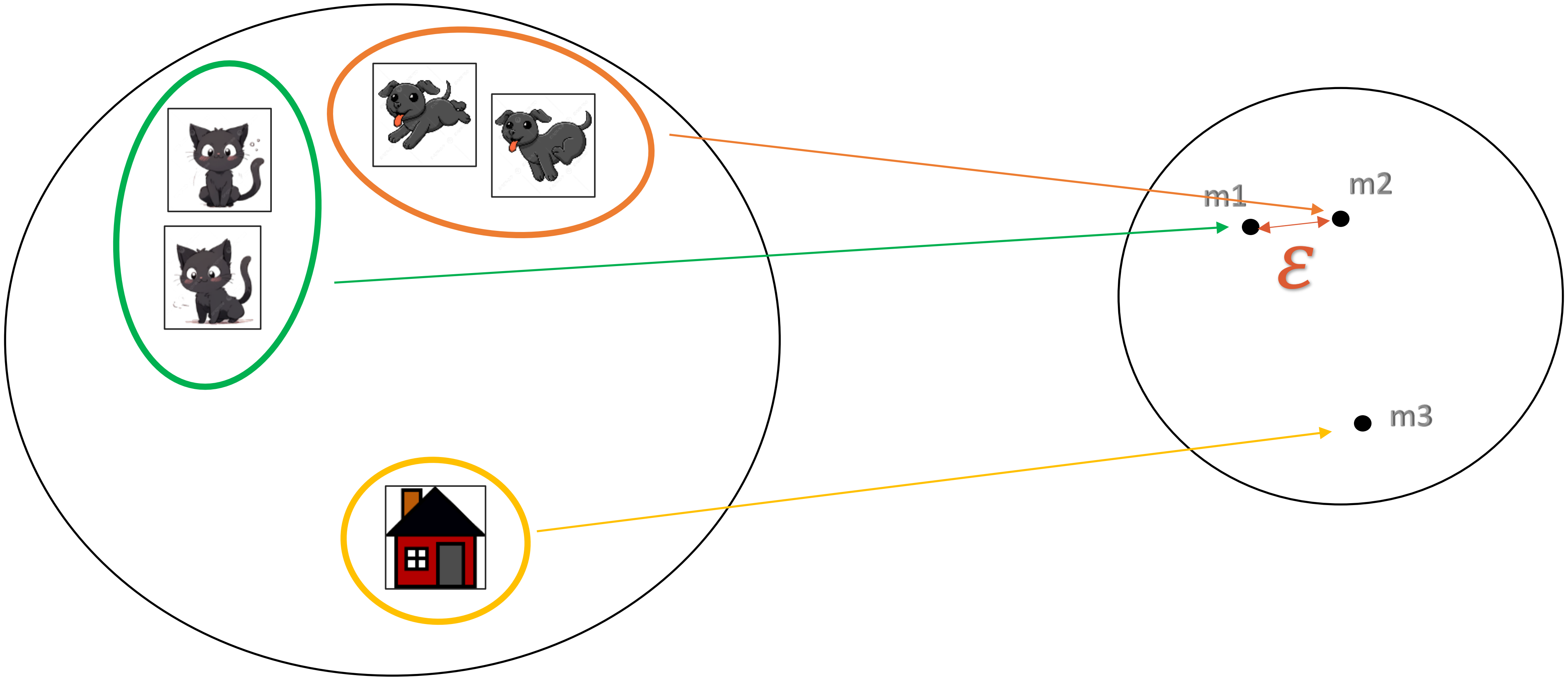
Spatial meaningfulness

Semantic consistency requires hard equality of messages

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

Definition ignores distances between messages

Ideally, we want: objects mapped to **similar messages** should be similar



m1
m2
ε
m3

Spatial meaningfulness

To consider distance between messages, define:

A communication protocol S_θ is **$\boldsymbol{\varepsilon_0}$-spatially meaningful** if $\forall 0 < \varepsilon \leq \varepsilon_0$

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

Spatial meaningfulness

To consider distance between messages, define:

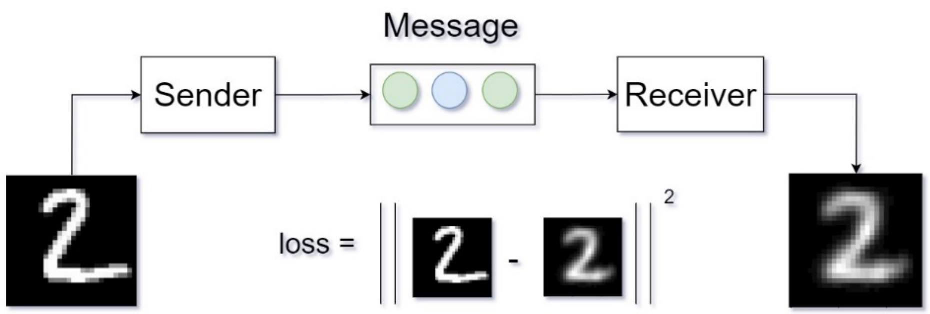
A communication protocol S_θ is **$\boldsymbol{\varepsilon_0}$-spatially meaningful** if $\forall 0 < \varepsilon \leq \varepsilon_0$

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

Need assumptions on receiver:

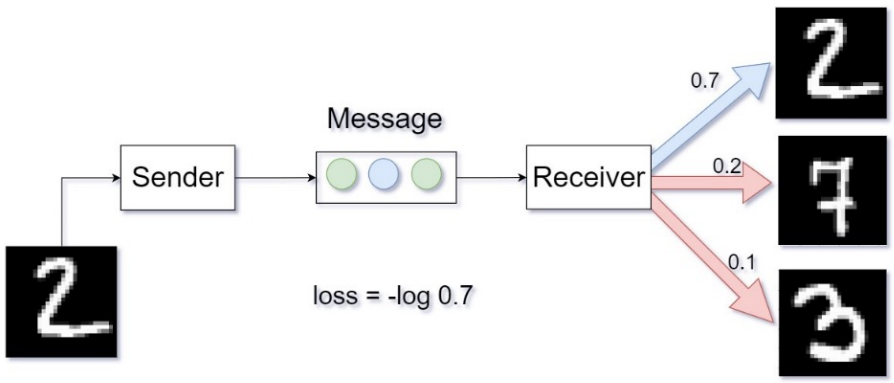
- **simple**: $\left\|R_\varphi(x_1) - R_\varphi(x_2)\right\| \leq k \cdot \|x_1 - x_2\|$
- **non-degenerate**: better than any constant receiver

Reconstruction



Theorem: Assuming receiver Φ is unrestricted, if receiver $\mathrm{R}_\varphi \in \Phi$ is (I, M)-simple and non-degenerate, then every S_{θ^*} that is conditionally optimal for R_φ is **spatially meaningful** with $\varepsilon_0 = \min_{m_1 \neq m_2} \|m_1 - m_2\|$.

Discrimination



Theorem: There exists a game, receiver R_φ and sender S_θ such that Θ is unrestricted, R_φ is (I, M)-simple and non-degenerate, S_θ is conditionally optimal matching R_φ, and S_θ is not spatially meaningful.

Back to reality

- Limited agents, with optimization problems
- No oracle natural language concepts
- No examples of emergent messages and **parallel** natural concepts

How can we decipher emergent communications?

A Theory of Unsupervised Translation Motivated by Understanding Animal Communication

Shafi Goldwasser*
UC Berkeley & Project CETI
shafi.goldwasser@berkeley.edu

David F. Gruber*
City University of New York & Project CETI
david@projectceti.org

Adam Tauman Kalai*
Microsoft Research & Project CETI
adam@kal.ai

Orr Paradise*
UC Berkeley & Project CETI
orrp@eecs.berkeley.edu

"[…] unsupervised translation of animal communication may be feasible if the communication system is sufficiently complex."

Unsupervised MT of emergent communication

1. Train agents to play a game
2. Collect many emergent messages
3. Separately, collect many English texts
   - From same domain
   - But not parallel
4. Train unsupervised machine translation from messages to English

Unsupervised MT of emergent communication

1. Train agents to play a game
2. Collect many emergent messages
3. Separately, collect many English texts
   - From same domain
   - But not parallel
4. Train unsupervised machine translation from messages to English

Experiment: agents describe images from MS COCO

Two people are standing in front of a bus

A train is traveling down the tracks near a platform

A bunch of food that are on a plate

A child sitting on a bed with a stuffed animal

Quantitative evaluation

Model	Category	Supercategory	Random	Inter-category	Baseline
Novelty (%)	58.74 ± 7.81	**70.00** ± 1.68	60.54 ± 4.25	57.36 ± 5.83	100.0
BLEU Score	7.41 ± 0.47	6.08 ± 0.31	6.85 ± 0.34	**9.21** ± 0.45	0.071
BERTScore	**0.734** ± 0.001	0.730 ± 0.001	0.729 ± 0.001	0.730 ± 0.001	0.543
METEOR Score	0.295 ± 0.06	0.276 ± 0.06	0.289 ± 0.06	**0.310** ± 0.07	0.115
ROUGE-L	0.361 ± 0.001	0.343 ± 0.006	0.352 ± 0.003	**0.370** ± 0.002	0.173
Jaro Similarity	0.678 ± 0.02	0.673 ± 0.02	0.676 ± 0.02	**0.682** ± 0.02	0.601
CLIP Score	0.180 ± 0.018	0.176 ± 0.019	0.183 ± 0.020	**0.191** ± 0.019	0.151
TTR (%)	0.42 ± 0.05	**0.71** ± 0.14	0.58 ± 0.11	0.59 ± 0.15	0.19

Quantitative evaluation

Model	Category	Supercategory	Random	Inter-category	Baseline
Novelty (%)	58.74 ± 7.81	**70.00** ± 1.68	60.54 ± 4.25	57.36 ± 5.83	100.0
BLEU Score	7.41 ± 0.47	6.08 ± 0.31	6.85 ± 0.34	**9.21** ± 0.45	0.071
BERTScore	**0.734** ± 0.001	0.730 ± 0.001	0.729 ± 0.001	0.730 ± 0.001	0.543
METEOR Score	0.295 ± 0.06	0.276 ± 0.06	0.289 ± 0.06	**0.310** ± 0.07	0.115
ROUGE-L	0.361 ± 0.001	0.343 ± 0.006	0.352 ± 0.003	**0.370** ± 0.002	0.173
Jaro Similarity	0.678 ± 0.02	0.673 ± 0.02	0.676 ± 0.02	**0.682** ± 0.02	0.601
CLIP Score	0.180 ± 0.018	0.176 ± 0.019	0.183 ± 0.020	**0.191** ± 0.019	0.151
TTR (%)	0.42 ± 0.05	**0.71** ± 0.14	0.58 ± 0.11	0.59 ± 0.15	0.19

Modest translation quality

Seems to capture main theme of the image, but **not details**

Next step: how does game **complexity** affect language richness

Contributions

- Emergent communication: playground for deciphering “alien” language
- Discrete codebook enables interpretable communication
- Theory: game complexity affects “naturalness” of emergent language
- Unsupervised translation: initial positive signs
  - ➢Time to try on animal communication?

Collaborators
Boaz Carmeli, Rom Meir,
Rotem Ben-Zion,
Ido Levy, Orr Paradise