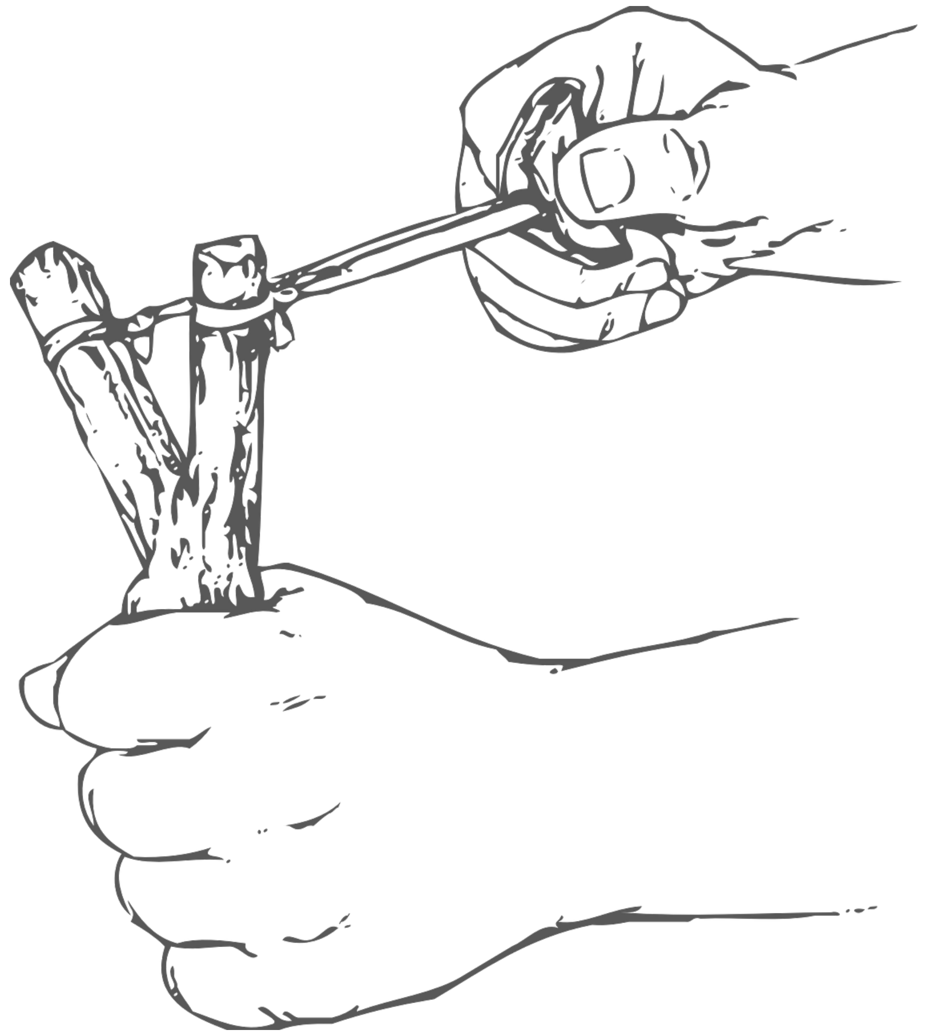
Neural·Pragmatic
Natural
Language
Generation

N·P
NLG

Learning goals

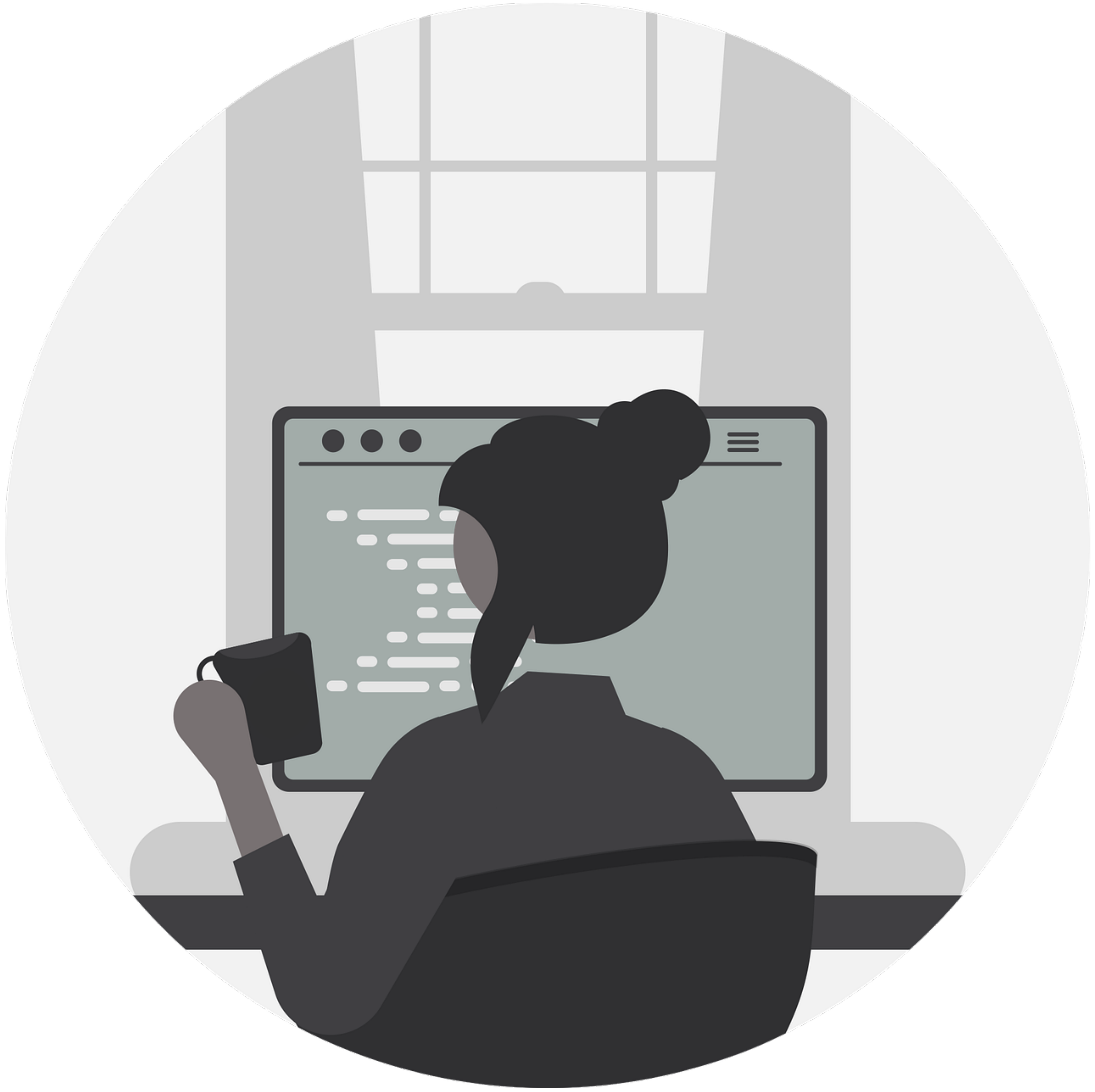
1. become oriented in the landscape of pragmatic neural NLG
2. understand different ways in which RSA(-like) ideas can be applied in NLG:
   a. during training
   b. during inference



organizational remarks

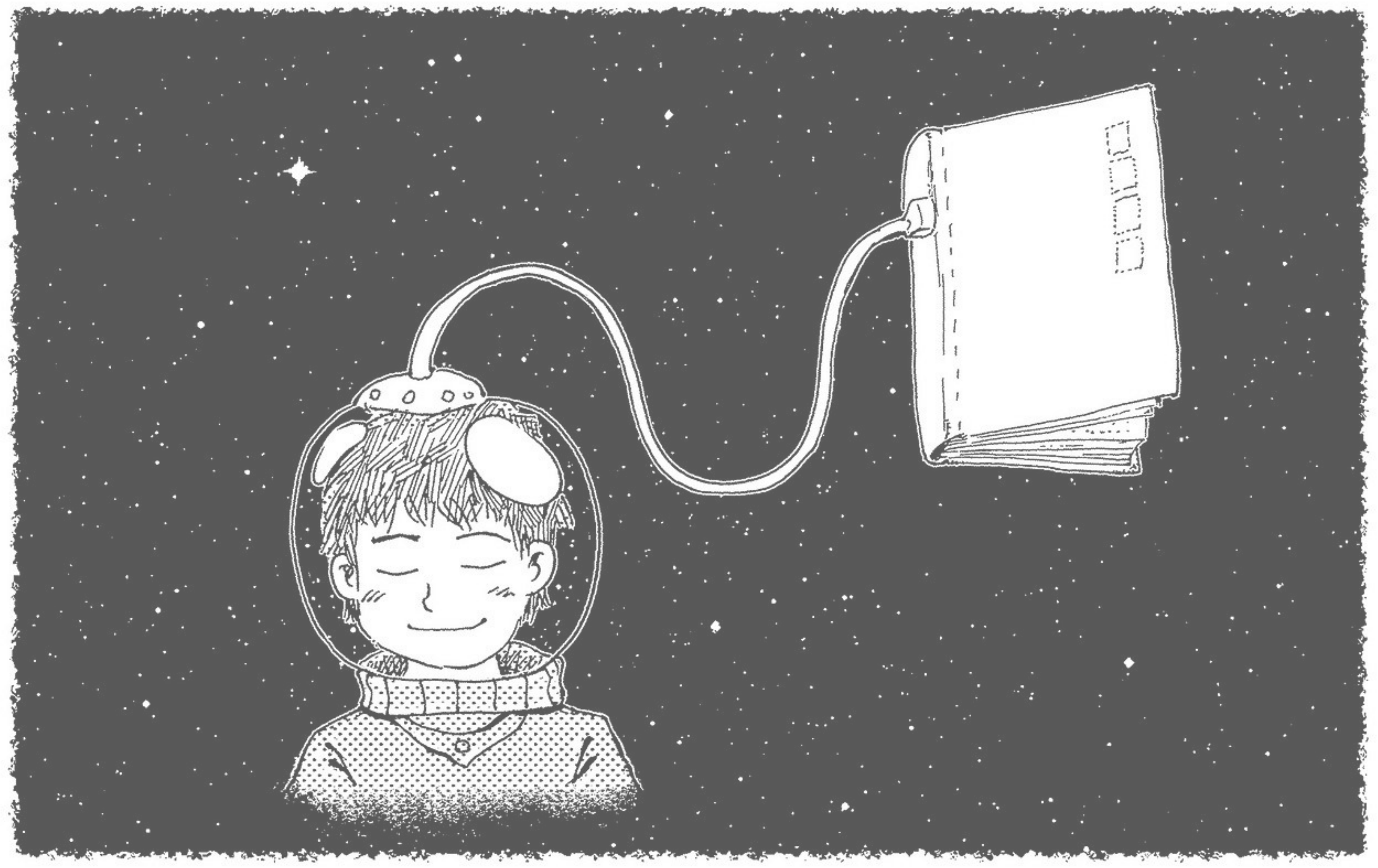
Course projects

- work in groups (2-3 people are ideal)
  - single-person projects are okay but need motivation & permission
  - problems in the group discussed w/ lecturer before escalation
  - there will be one grade for the whole group
- outcome of the project
  - structured, documented, self-contained repository w/ all materials
  - highly accessible (reproducible, commented …) code
  - short research paper (PDF) explaining what was done, how this relates the to literature, why it was done and what was achieved or found
- content & scope
  - critical conceptual / mathematical work (even w/o any code) is welcome
  - typical project will aim to reproduce key results from a single paper
  - ambitious projects can shine by additionally:
    - extending or combining existing analyses
    - critically discussing existing analyses (in the light of the literature or project results)
    - conceptually motivated exploration of novel models, different data sets, other evaluation measures …



How to read a research paper

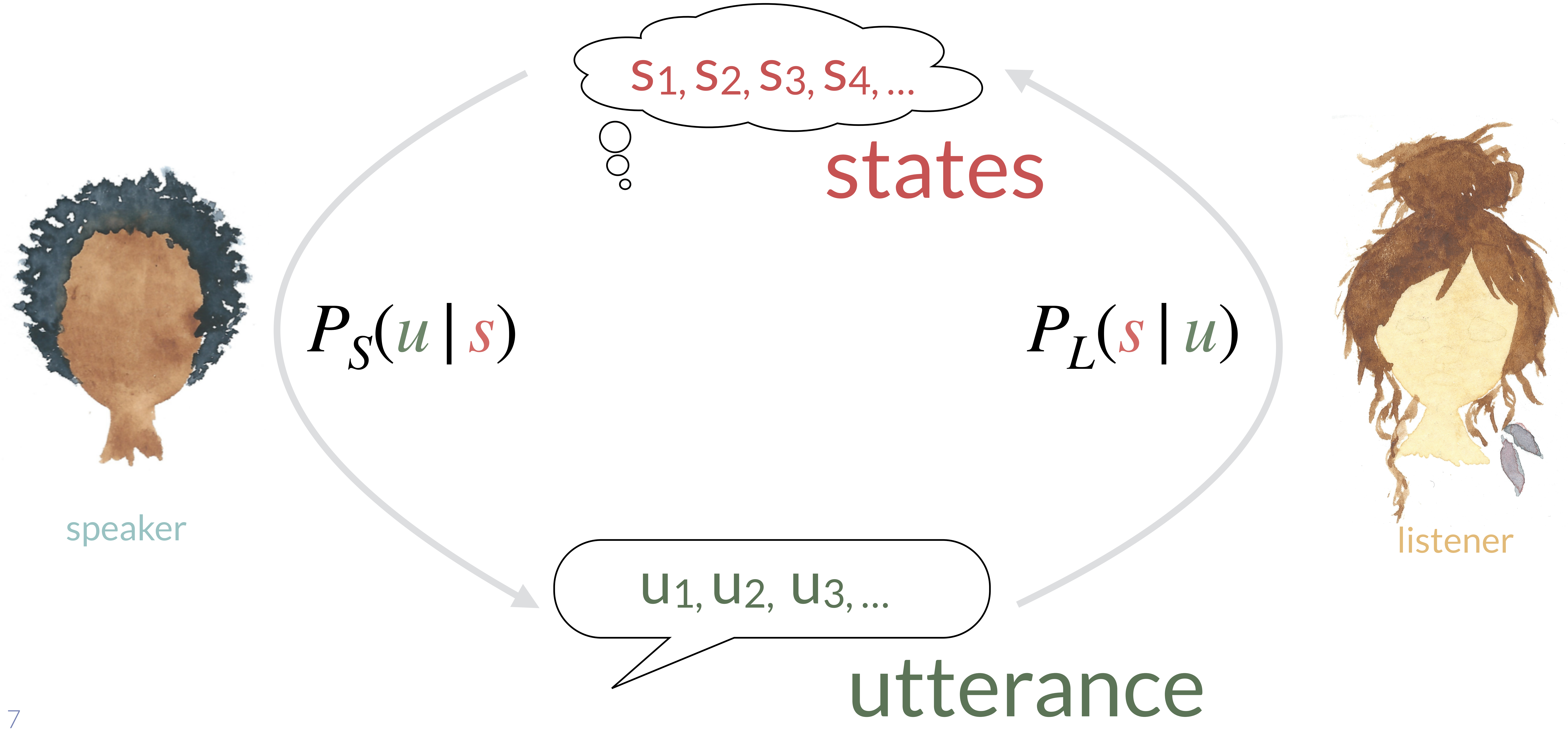
- identify key innovation / argument / point of the paper
  - how novel or important is this?
- track what you like and dislike
  - e.g., what's well explained, what's incomprehensible?
  - how can you incorporate what's good into your own repertoire?
  - how would *you* have done it differently?
- track what / how much you understand
  - what would I need in addition to understand more?
  - what don't I understand that I don't need to understand?
- take notes
  - organize and revisit your notes



RSA meets neural NLG

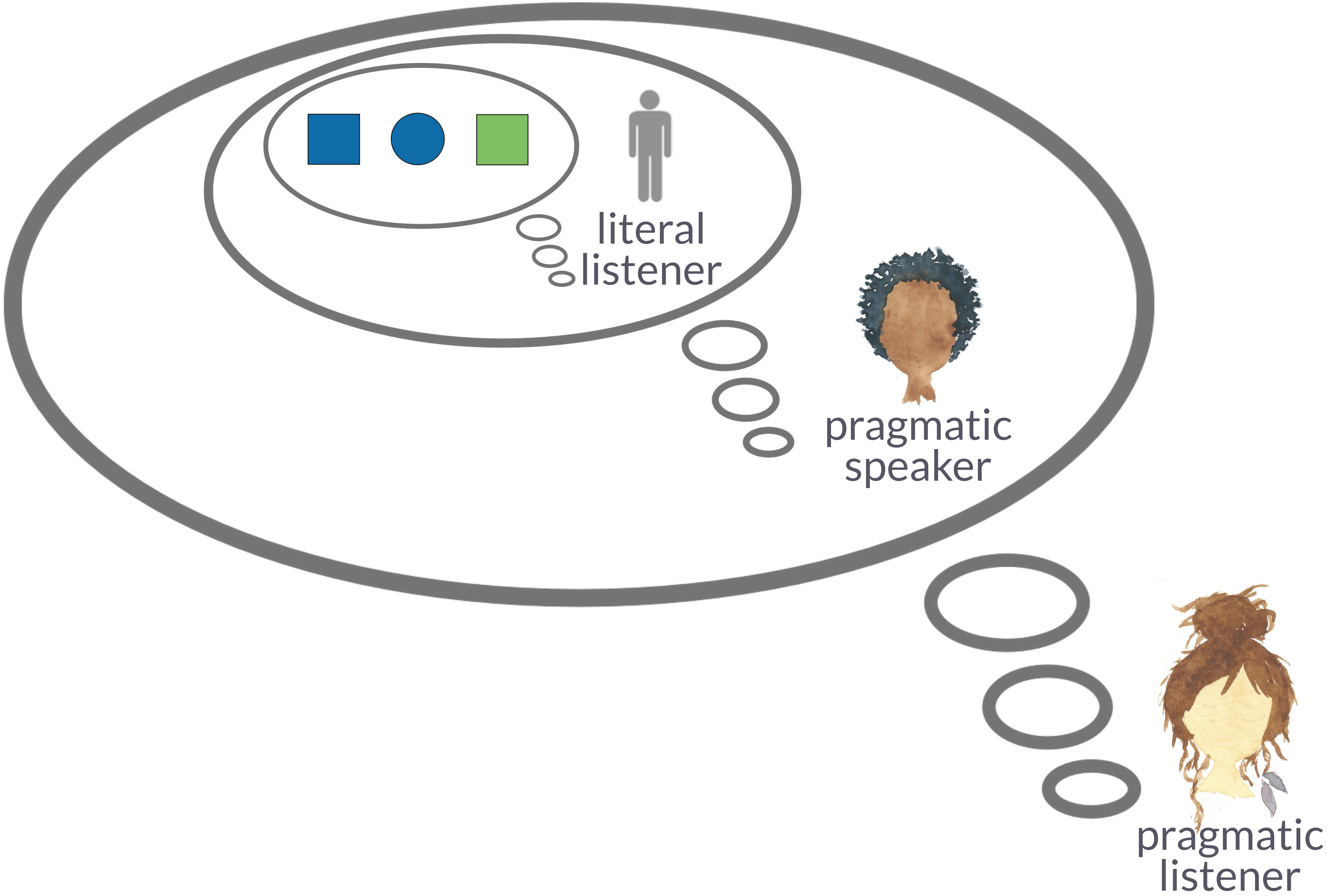
Pragmatic back-and-forth reasoning

speaker and listener reason about each other's behavior in a share context



Grounding pragmatic reasoning

in a (dummy) literal listener



RSA-style
literal listener grounding

literal L0-listener

pragmatic L1-speaker

hyper-pragmatic L2-listener

“Inverse-RSA”
literal speaker grounding

literal L0-speaker

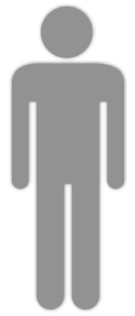
pragmatic L1-listener

hyper-pragmatic L2-speaker

Rabin (1990), Franke & Jäger (2014)

“standard RSA”

literal listener grounding



literal L0-listener

$$P_{L_0}(s \mid u) \propto P(s) \;\; \mathfrak{L}(s, u)$$

pragmatic L1-speaker

$$P_{S_1}(u \mid s) = \mathrm{SM}_\alpha \left(\log P_{L_0}(s \mid u) - \mathrm{C}(u) \right)$$

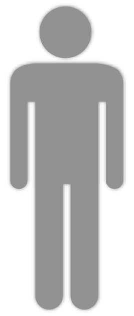
hyper-pragmatic L2-listener

$$P_{L_2}(s \mid u) \propto P(s) \;\; P_{S_1}(u \mid s)$$

⋮

“inverse RSA”

literal speaker grounding



literal L0-speaker

$$P_{S_0}(u \mid s) \propto P(u) \;\; \mathfrak{L}(u, s)$$

pragmatic L1-listener

$$P_{L_1}(s \mid u) \propto P(s) \;\; P_{S_0}(u \mid s)$$

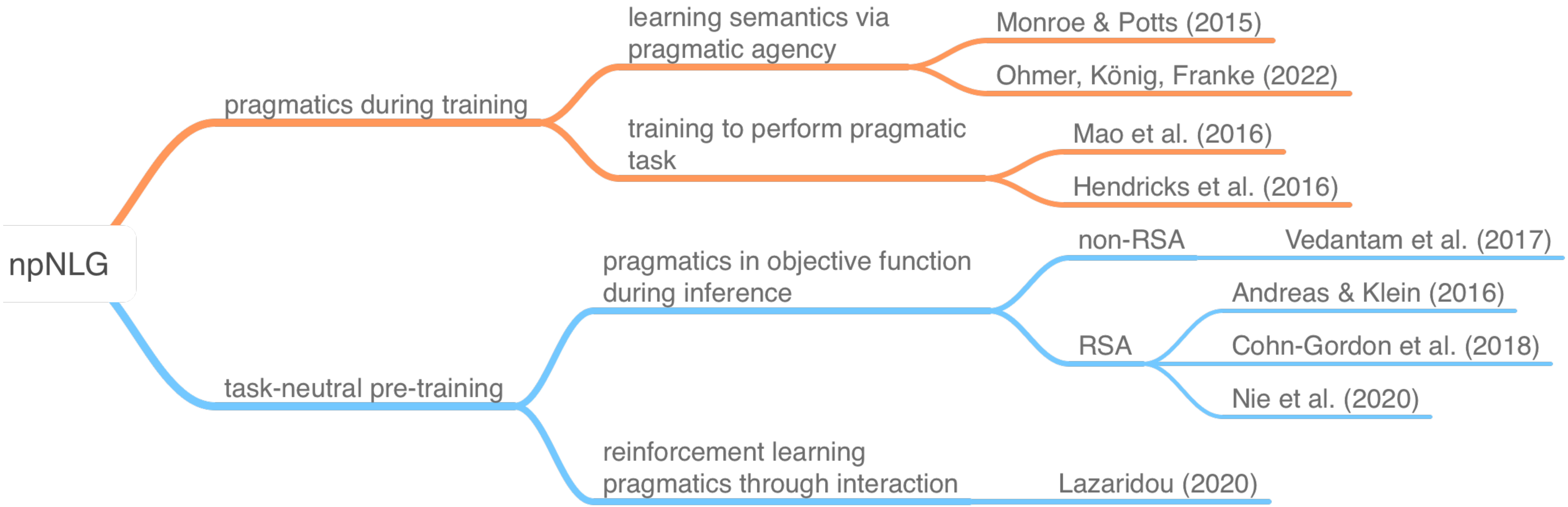
hyper-pragmatic L2-speaker

$$P_{S_2}(u \mid s) = \mathrm{SM}_\alpha \left(\log P_{L_1}(s \mid u) - \mathrm{C}(u) \right)$$

⋮

Overview

different kinds of npNLG approaches



Learning in the RSA model

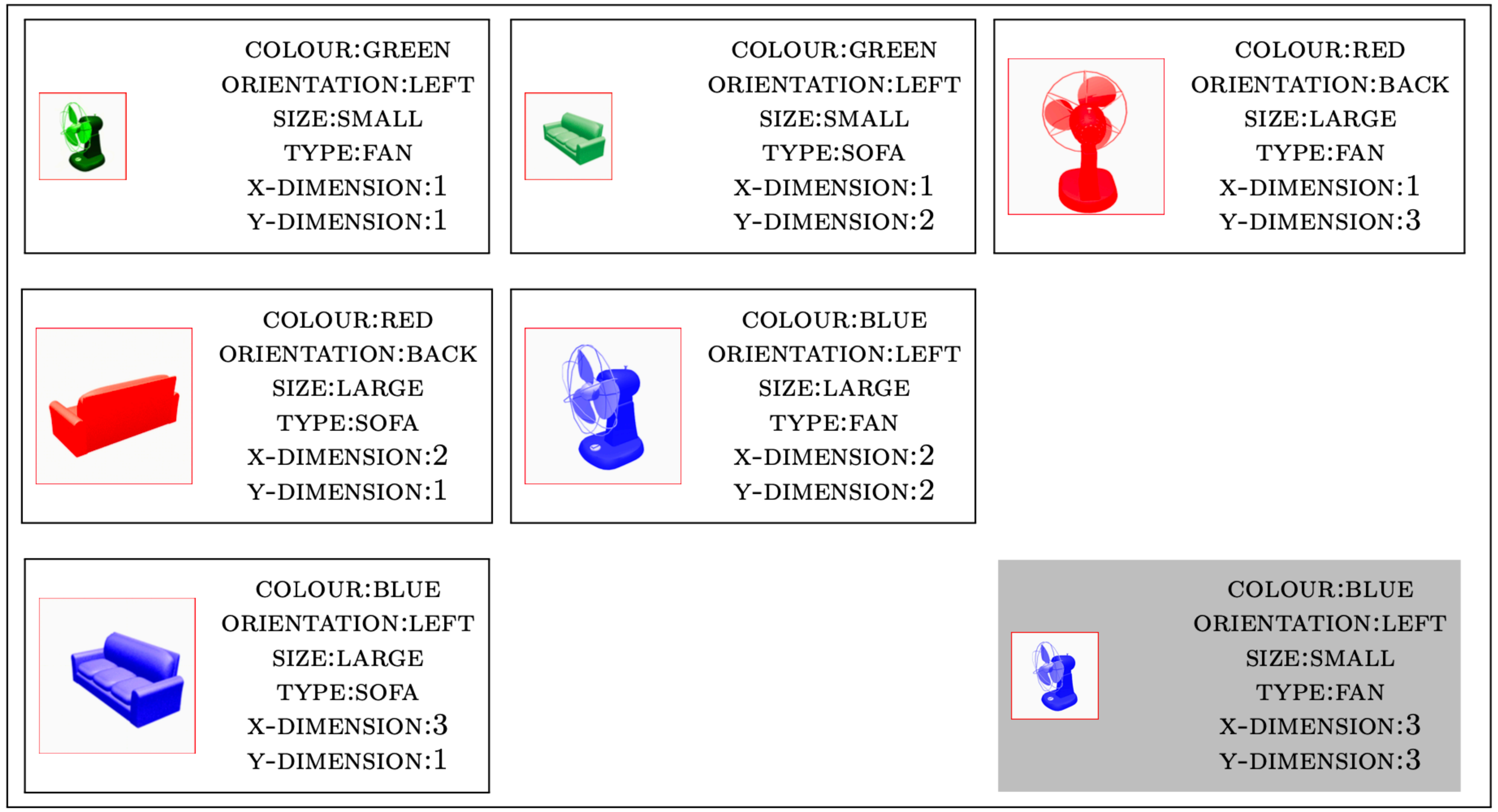
Monroe & Potts (2015), Proc. of Amsterdam Colloquium

Learning in the RSA model

data & modeling set-up

- **goal:** use empirical data to infer semantic meaning that optimizes performance of a speaker model (literal or pragmatic)
- data from TUNA corpus
  - human referential descriptions
  - annotated discrete features of objects
- literal meanings are learned from corpus data
  - $\mathfrak{L}(s, u, c) = \theta^T \varphi(s, u, c)$, where
    - θ^T is a linear mapping
    - $\varphi(s, u, c)$ is a feature representation function
- inverse RSA architecture
  - $P_{S_0}(u \mid s, c) = SM_\alpha\left(\mathfrak{L}(s, u, c)\right)$
  - $P_{L_1}(s \mid u, c) \propto P_{S_0}(u \mid s, c)$
  - $P_{S_2}(u \mid s, c) = SM_\alpha\left(P_{L_1}(s \mid u, c)\right)$

example from the TUNA corpus



Utterance: "blue fan small"
Utterance attributes: [*colour:blue*]; [*size:small*]; [*type:fan*]

Monroe & Potts (2015)

Learning in the RSA model
evaluation & results

- evaluation metrics:
  - compare features selected by human & machine
  - **accuracy:** perfect match in all features
  - **dice score:** degree of overlap selected features
- models compared:
  - untrained RSA (just using features)
  - speaker models with learned semantics:
    - literal vs pragmatic speakers
    - based on different kinds of features:
      - basic features
      - additional information on human-like generation
- upshot & evaluation:
  - outperforms RSA (w/ predefined meanings)
  - trained S1 is best on aggregate data
  - **BUT:** requires a curated set of discrete features

results reported in the paper

	Furniture		People		All	
Model	Acc.	Dice	Acc.	Dice	Acc.	Dice
RSA s_0 (random true message)	1.0%	.475	0.6%	.125	1.7%	.314
RSA s_1	1.9%	.522	2.5%	.254	2.2%	.386
Learned S_0, basic feats.	16.0%	.779	9.4%	.697	12.9%	.741
Learned S_0, gen. feats. only	5.0%	.788	7.8%	.681	6.3%	.738
Learned S_0, basic + gen. feats.	**28.1%**	**.812**	17.8%	.730	***23.3%***	***.774***
Learned S_1, basic feats.	23.1%	.789	11.9%	.740	17.9%	.766
Learned S_1, gen. feats. only	17.4%	.740	1.9%	.712	10.3%	.727
Learned S_1, basic + gen. feats.	***27.6%***	.788	**22.5%**	**.764**	**25.3%**	**.777**

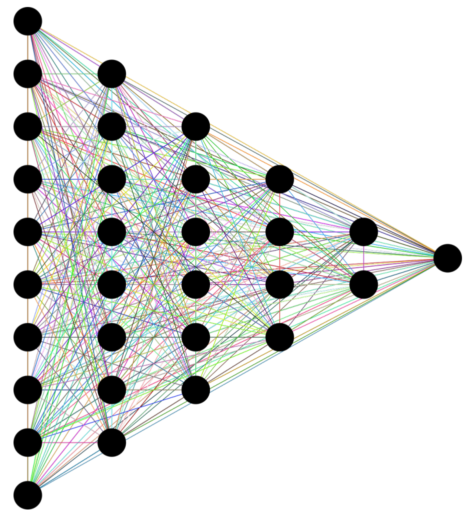
Monroe & Potts (2015)

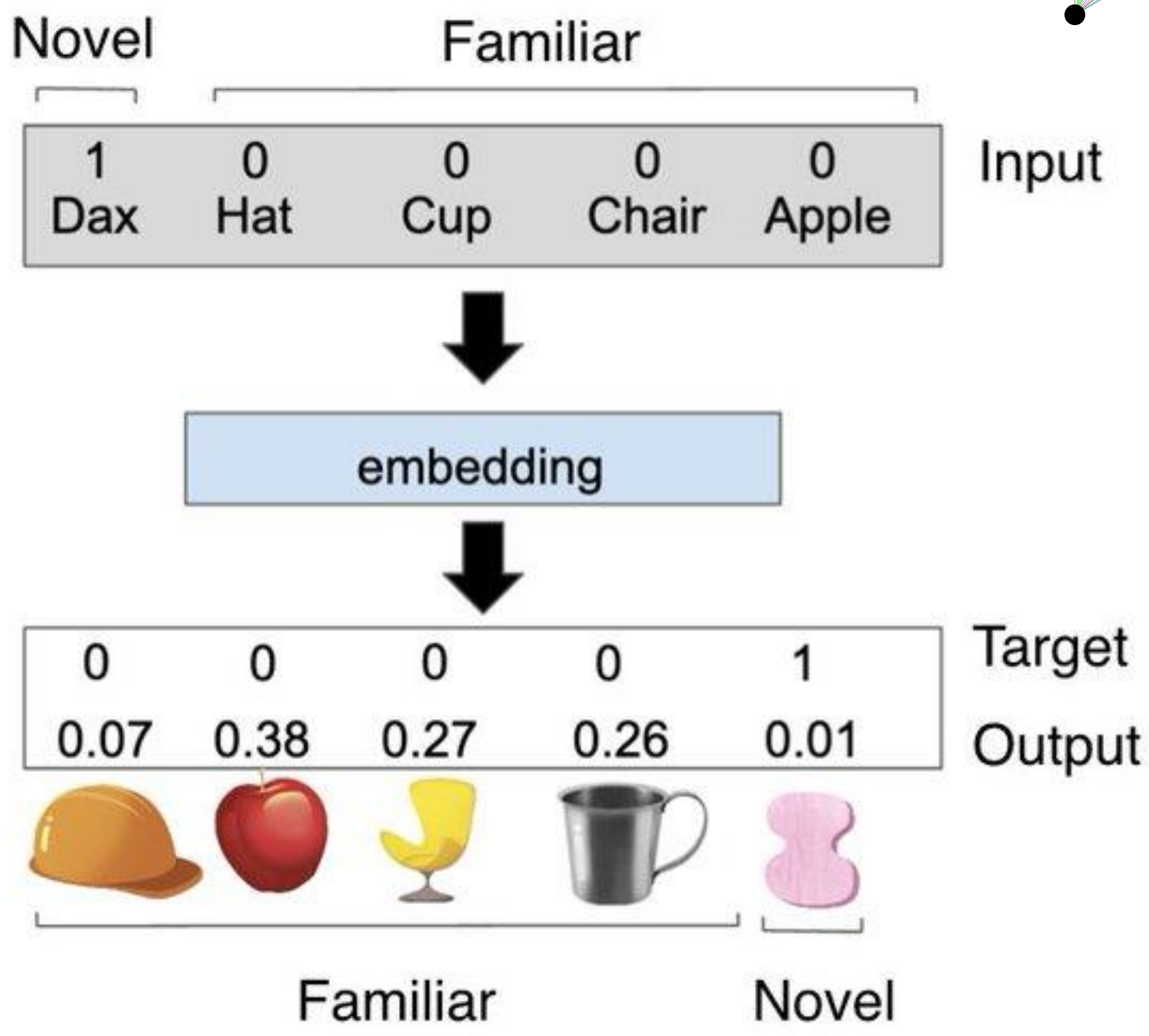
Pragmatic Reinforcement Learning

Ohmer, Franke & König (2021), Cognitive Science

Mutual exclusivity (ME) bias

Anti-ME bias in neural networks
Ghandi & Lake (2020, *arXiv*)

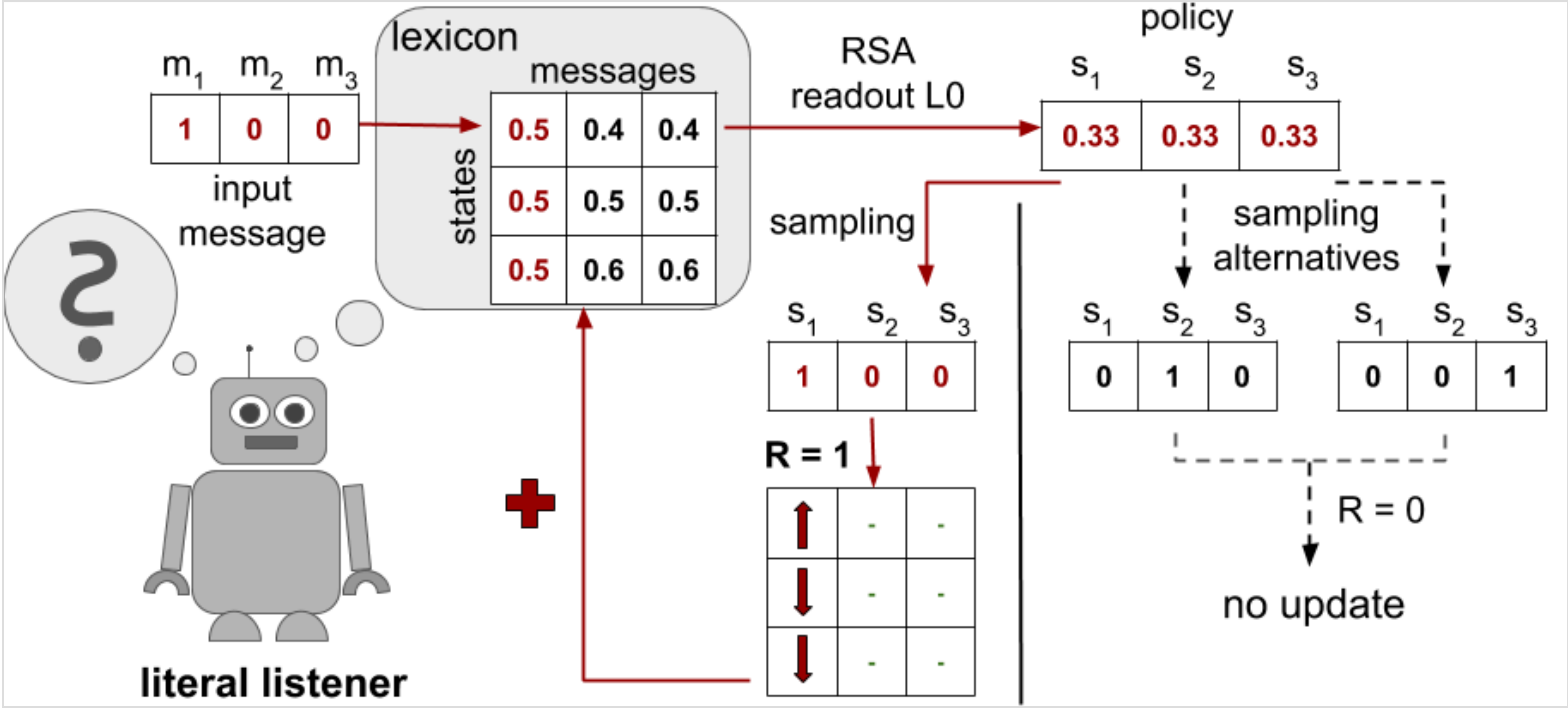




Gradient-based RL of semantic values

literal agents

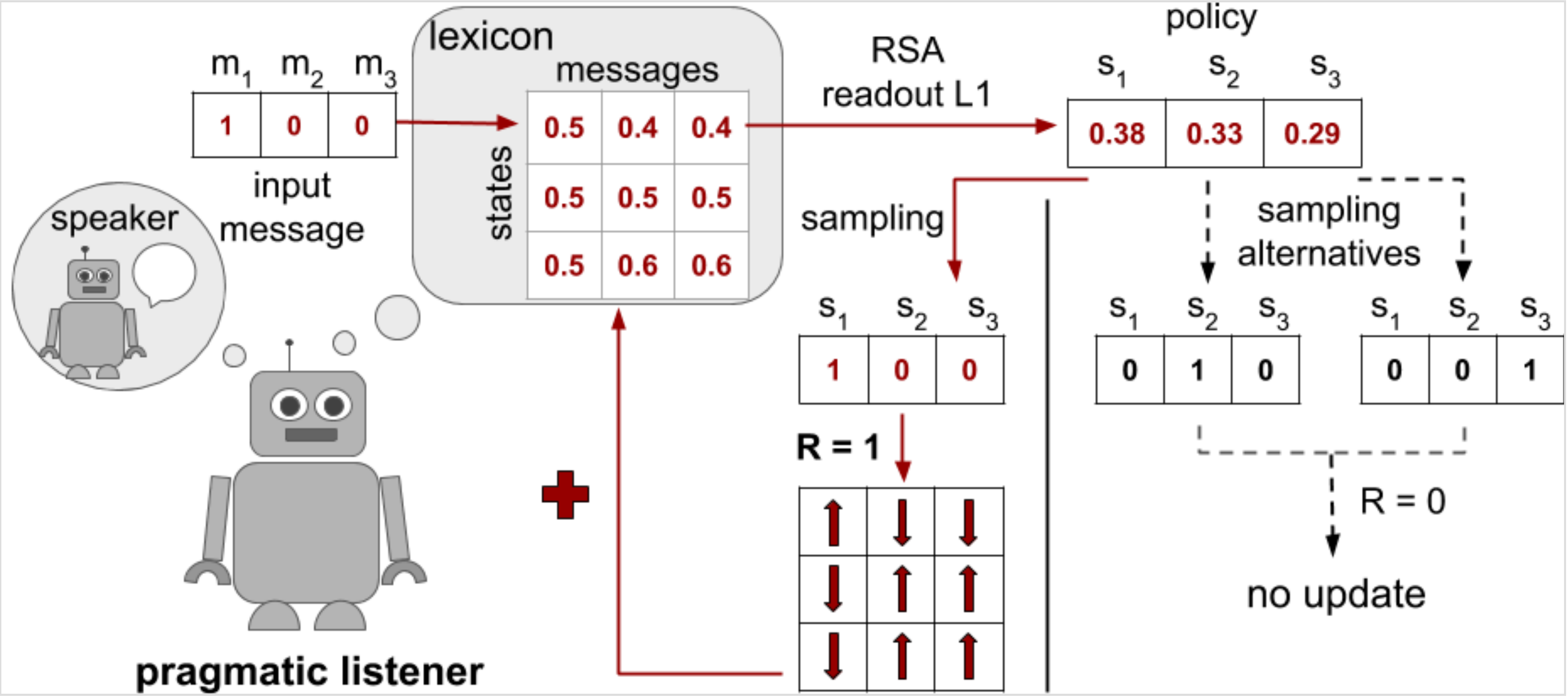
- agents update lexical meanings via RL
- policy defined by lexicon



Gradient-based RL of semantic values

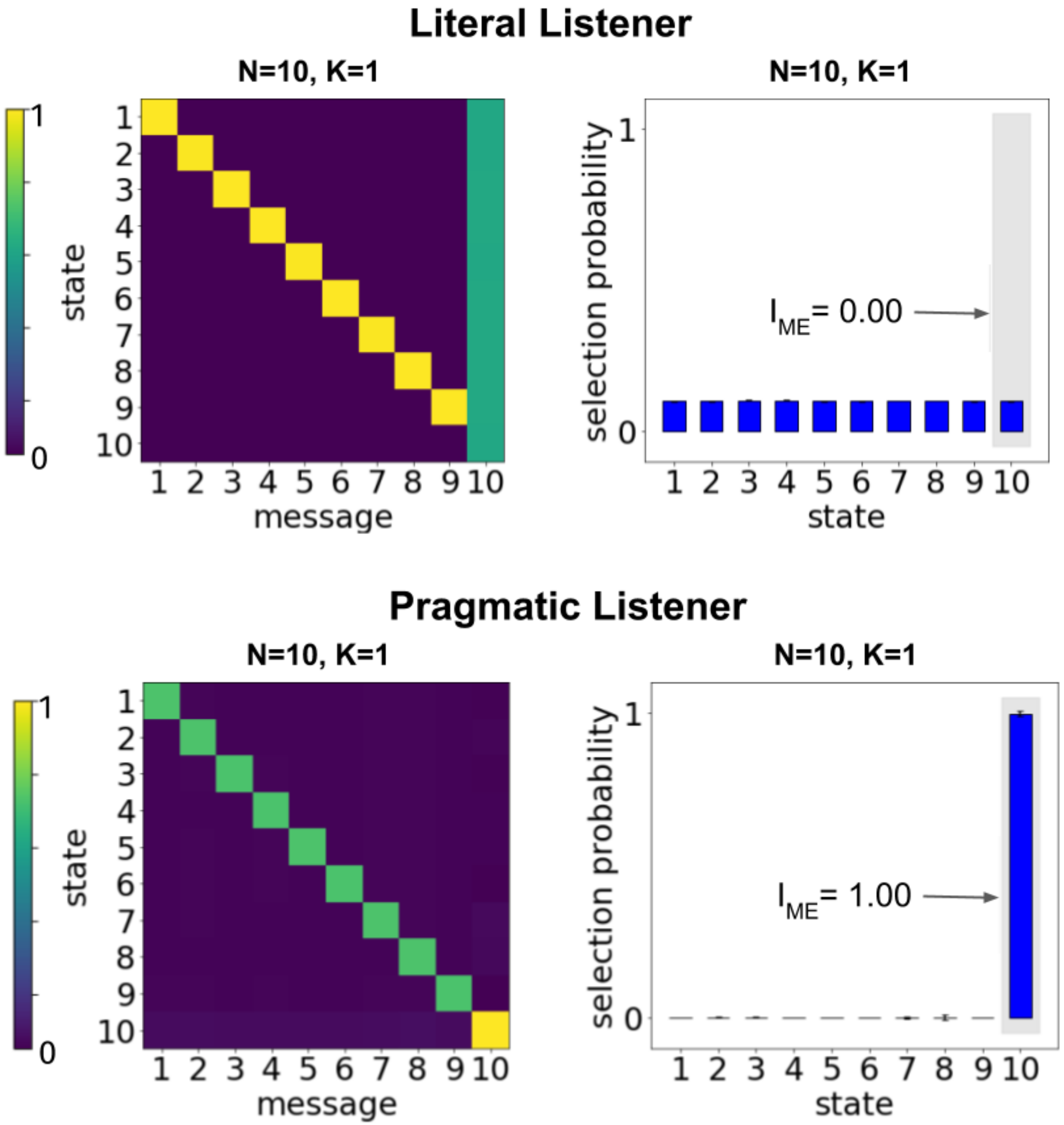
pragmatic agents

- agents update lexical meanings via RL
- policy defined by lexicon **& RSA**



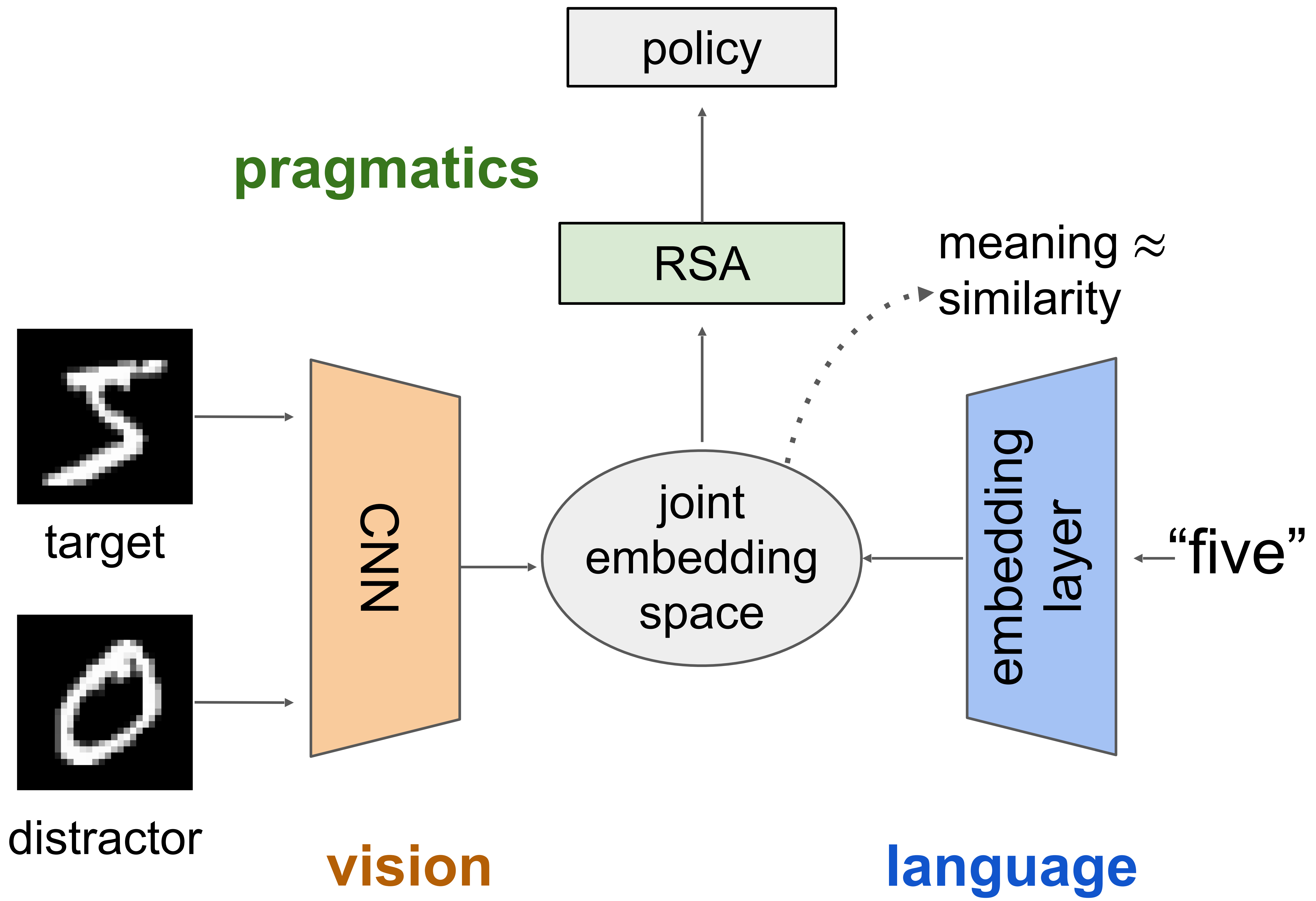
Simulation set-up & results

- set-up:
  - 10 states and messages matched 1-to-1
  - 9 pairs for training
  - 1 hold-out pair (index 10) for testing
- results:
  - lexical and behavioral ME bias for pragmatic agents, but not for literal agents
- extensions:
  - dynamically growing lexica
  - similarities to human word learning:
    - ME increases with vocabulary size
    - ME increases with exposure



Pragmatic RL in open-ended message & state spaces

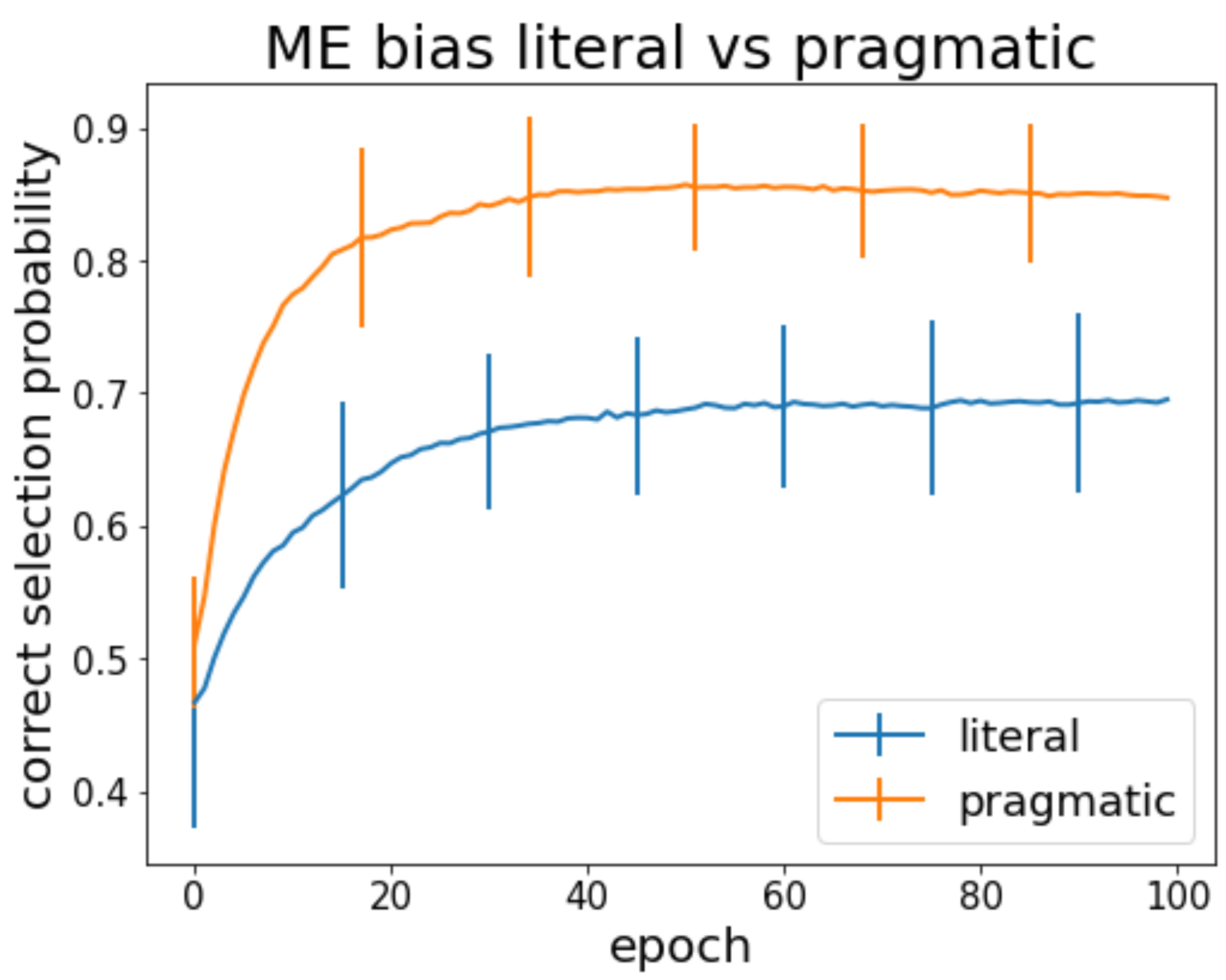
- image embedding
  $f: I \to [0;1]^n$
- message embedding
  $g: M \to [0;1]^n$
- semantic meaning:
  $\mathfrak{L}(s,m) = f(s) \cdot g(m)$

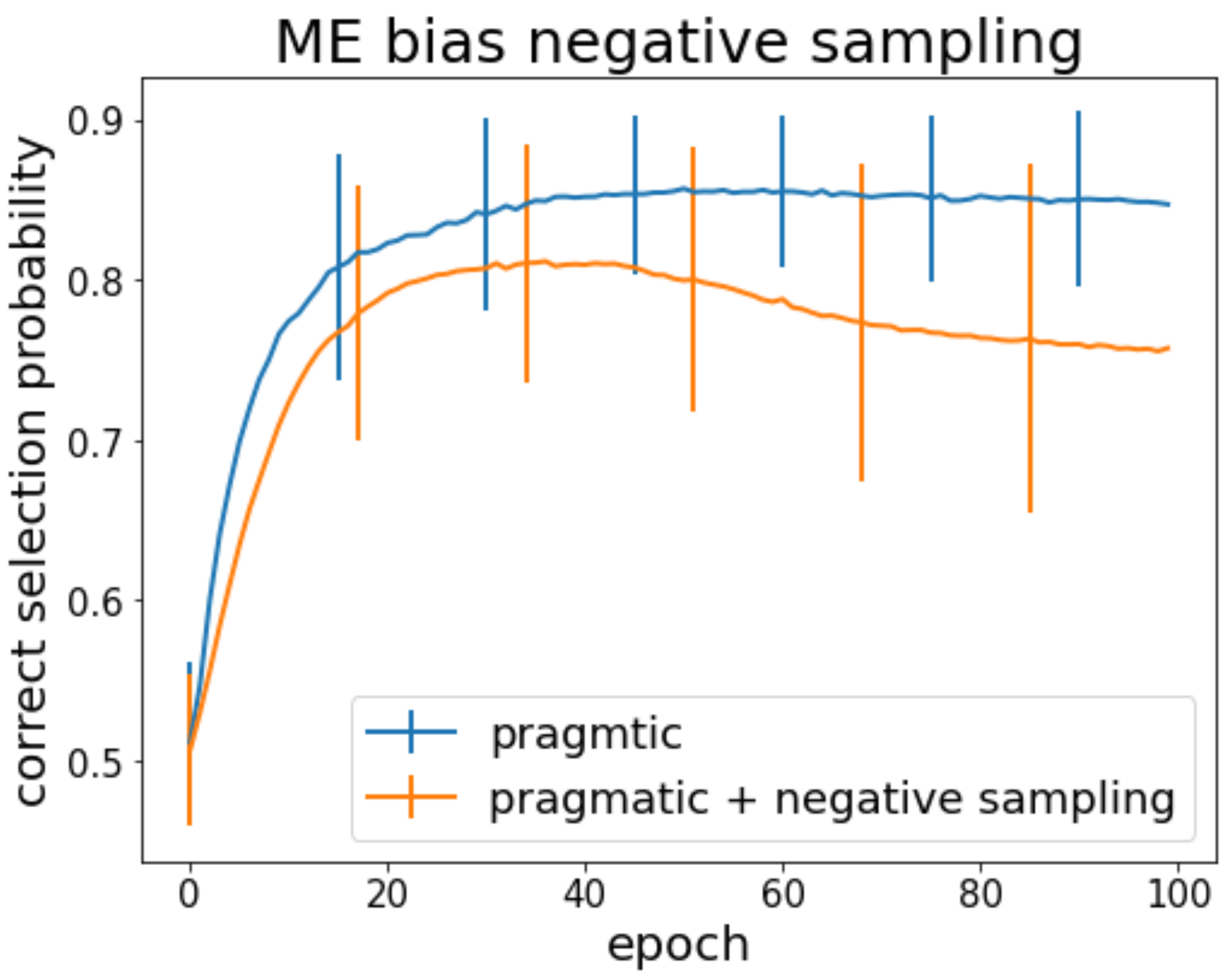


Simulation set-up & results

pragmatic RL w/ joint image-word embeddings

- set-up:
  - MNIST images as states
  - single embedding layer for single-word messages
  - one hold-out state/message
- results:
  - agents show behavioral ME bias
- negative sampling:
  - include non-matching image-word pairs during training marked as “negative examples”
    - Gulordava et al (2020); Vong & Lake (2022)
  - not required w/ pragmatic RL, even detrimental





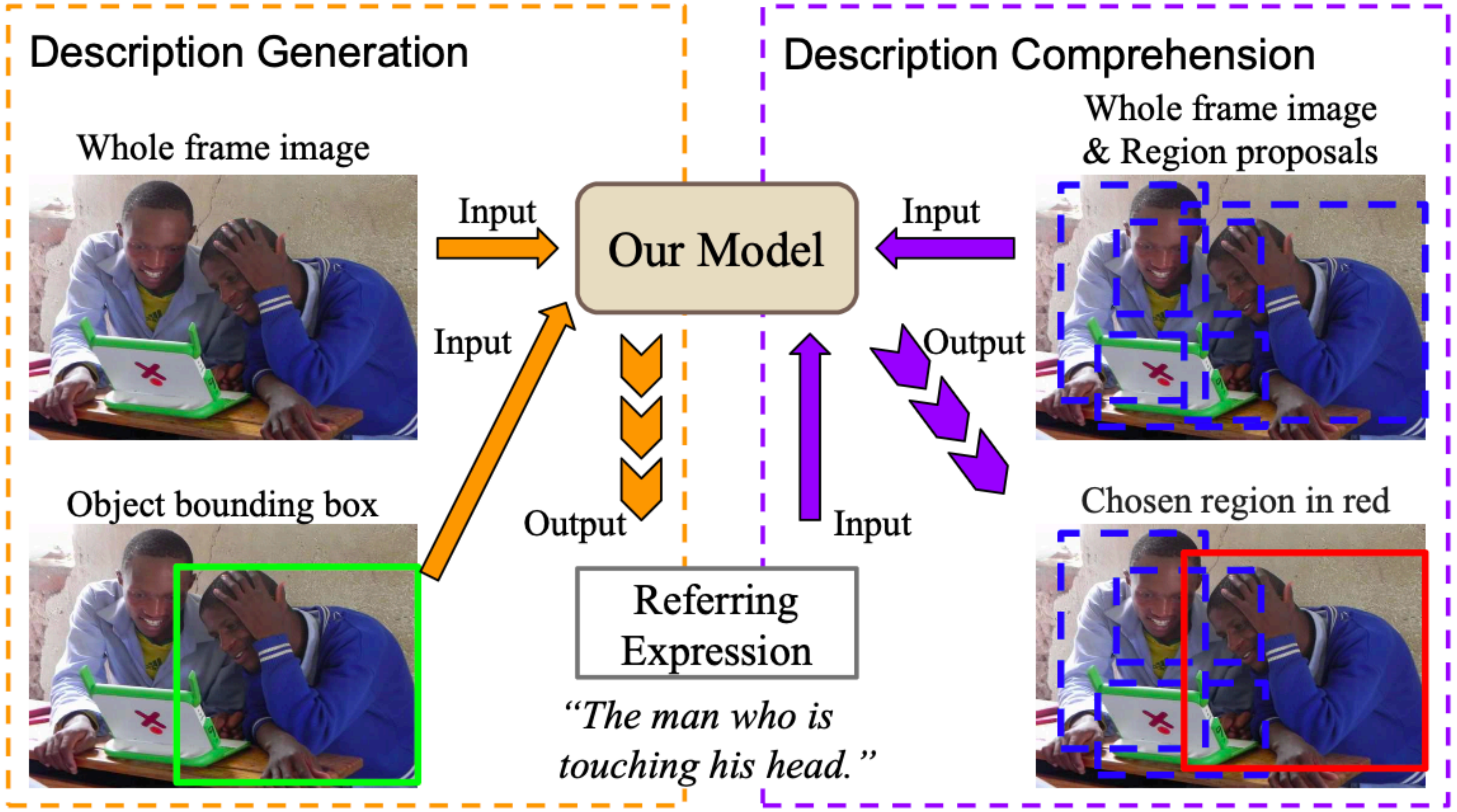
Generation and comprehension of unambiguous object descriptions

Mao et al. (2016), CVPR

Pragmatic object reference

learning context-discriminative object descriptions

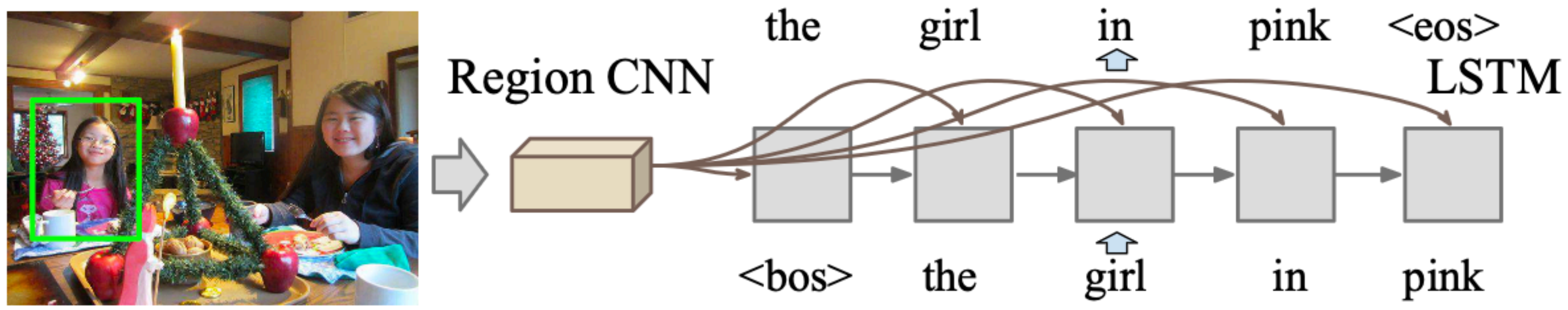
- **task:**
  - generate (unambiguous) referential description for a target object in an image
  - infer the intended referent object from a given description in an image
- **training set:**
  - Google Refexp data set
  - data points are triples: $\langle c, i, r \rangle$
    - caption
    - image
    - region (bounding box, represents objects)
- **approach:**
  - train S_0 and S_2 from "inverse RSA"



Mao et al. (2016)

Pragmatic object reference

system architecture



- literal speaker:
  - $P_{S_0}(c \mid i, r)$
  - trained as image captioner w/ objective function:

    $-\log P_{S_0}(c \mid i, r)$

- pragmatic listener:
  - $P_{L_1}(r \mid c, i) \propto P_{S_0}(c \mid i, r)$ [uniform priors]
  - implicit competitor set $R(i)$:
    - all objects in the picture
    - all objects of the same category
    - randomly generated bounding boxes

- pragmatic speaker:
  - $P_{S_2}(c \mid i, r) \propto P_{L_1}(r \mid c, i)$ [α = 1]
  - trained as image captioner w/ objective function:

    $-\log P_{L_1}(r \mid c, i)$ [max. mutual information]

Mao et al. (2016)

Pragmatic object reference

results

- human raters: percentage of generated descriptions that are at least as good as the description in the data set:
  - 15.9% for S_0
  - 20.4% for S_1

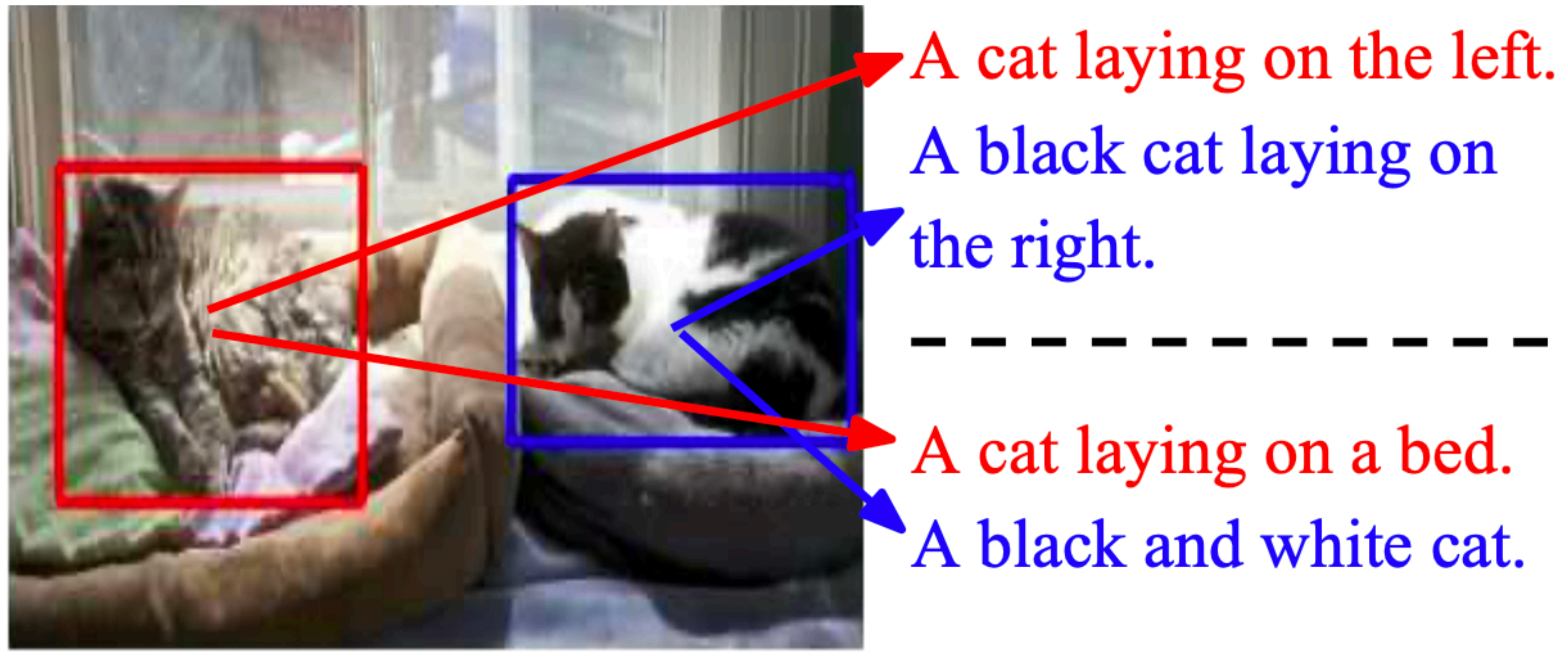
- accuracy of generated descriptions

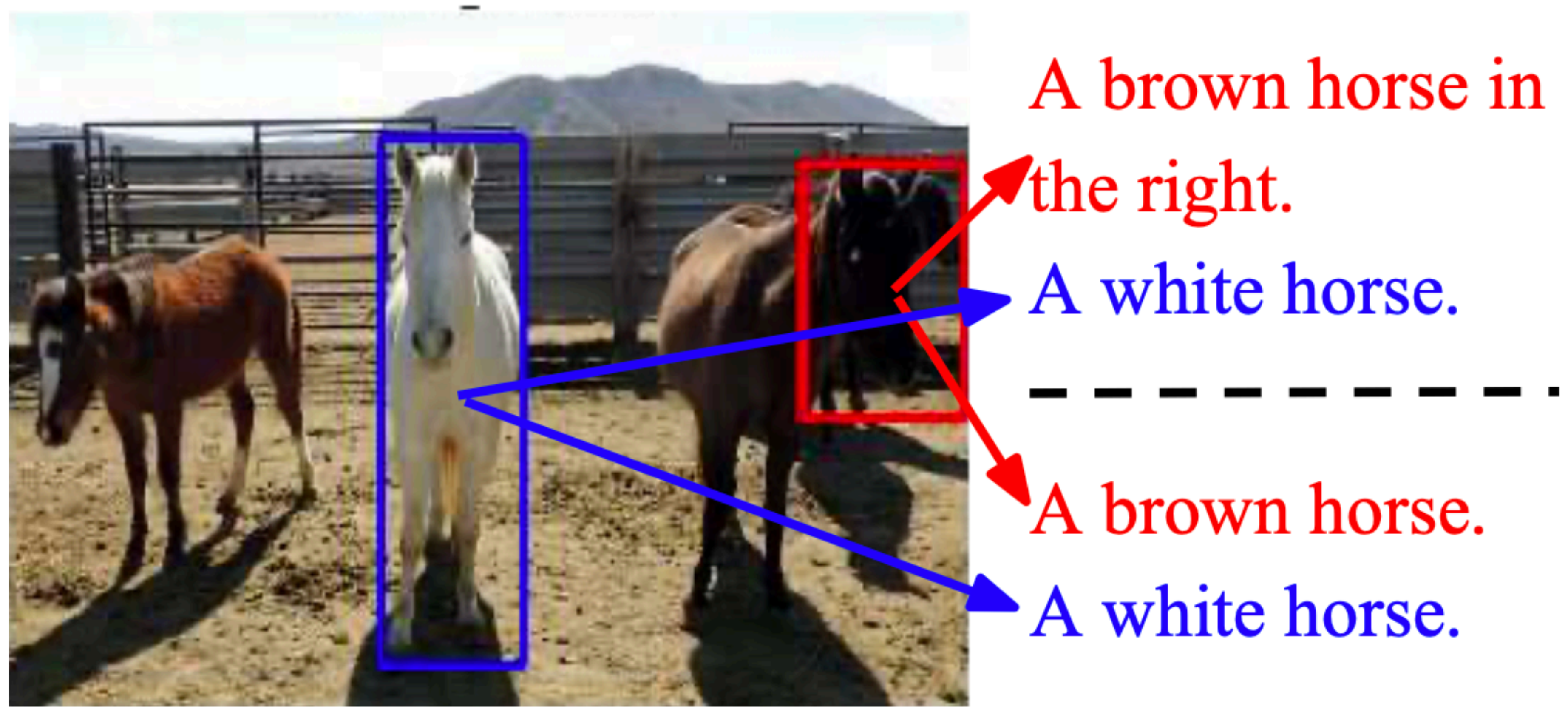
different competitor sets at test time

	Proposals / Descriptions	GT GEN	GT GT	Multibox GEN	Multibox GT
S_0	ML (baseline)	0.803	0.654	0.564	0.478
S_2	MMI-MM-easy-GT-neg	0.851	0.677	0.590	0.492
S_2	MMI-MM-hard-GT-neg	**0.857**	**0.699**	0.591	0.503
S_2	MMI-MM-multibox-neg	0.848	0.695	**0.604**	**0.511**
S_2	MMI-SoftMax	0.848	0.689	0.591	0.502

synthetic data

human data





Mao et al. (2016)

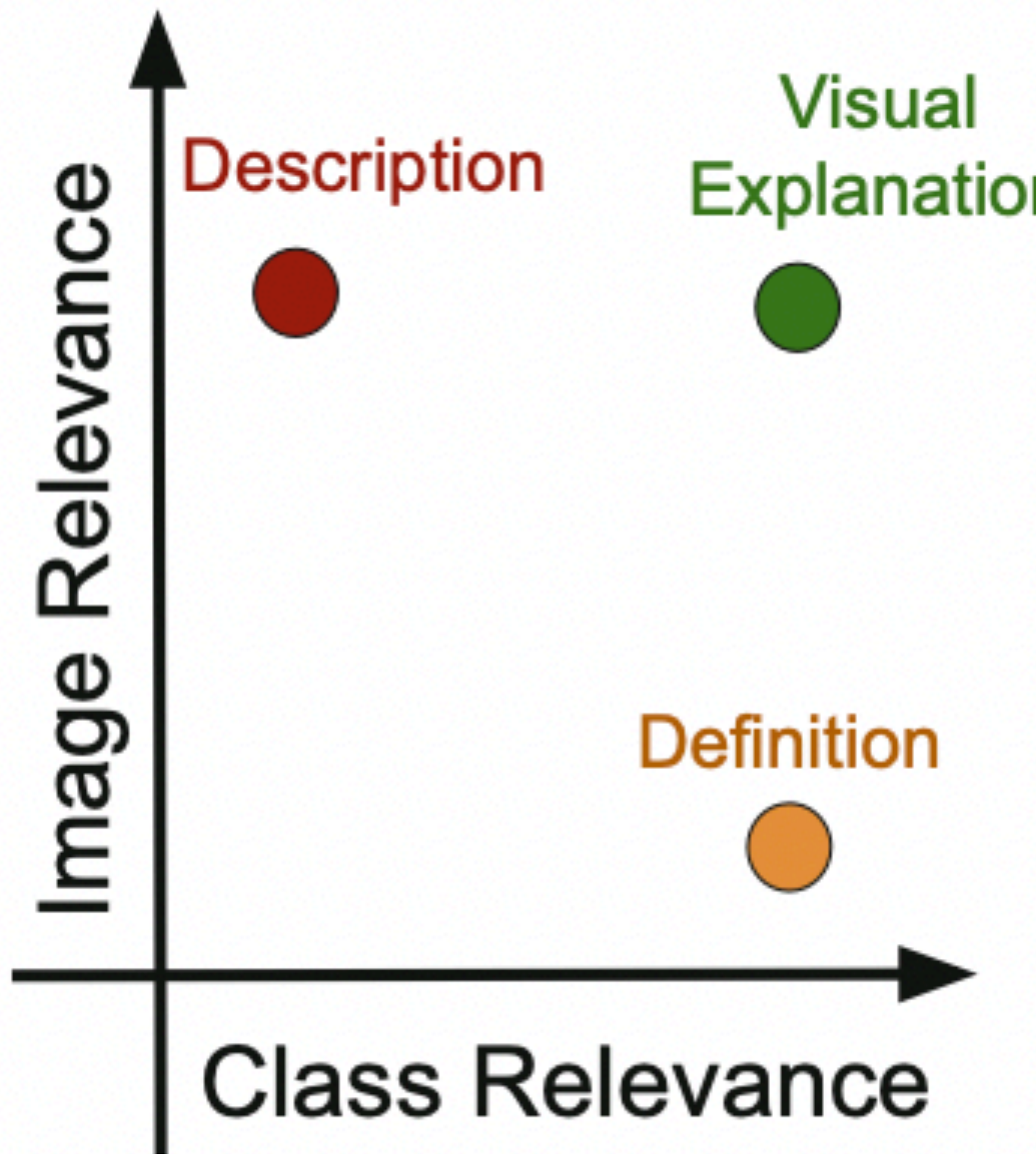
Generating visual explanations

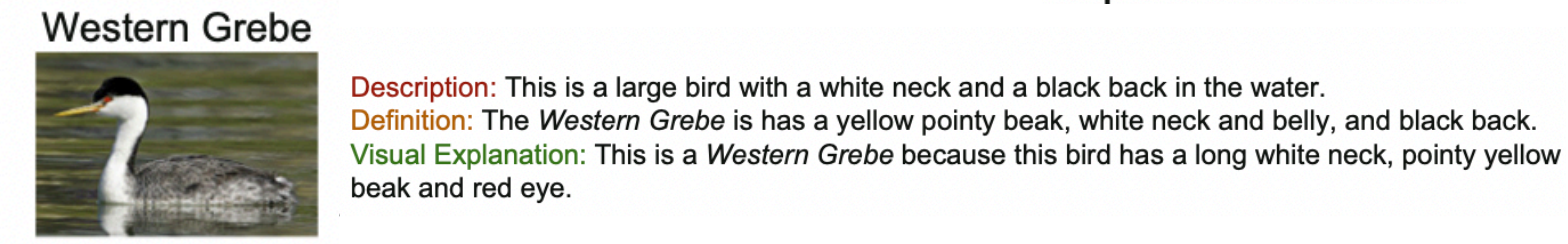
Hendricks et al. (2016), ECCV

Generating visual explanations

overview

- **goal:** produce caption for image i that justifies why i is an instance of given category C
- **data:** caption-image-category triples $\langle c, i, C \rangle$
  - CUB-justify data set
- approach:
  - S1-like agent, similar to Andreas & Klein (2016)
  - all pragmatics trained-in (like Mao et al. (2016)
  - loads of performance bells-&whistles

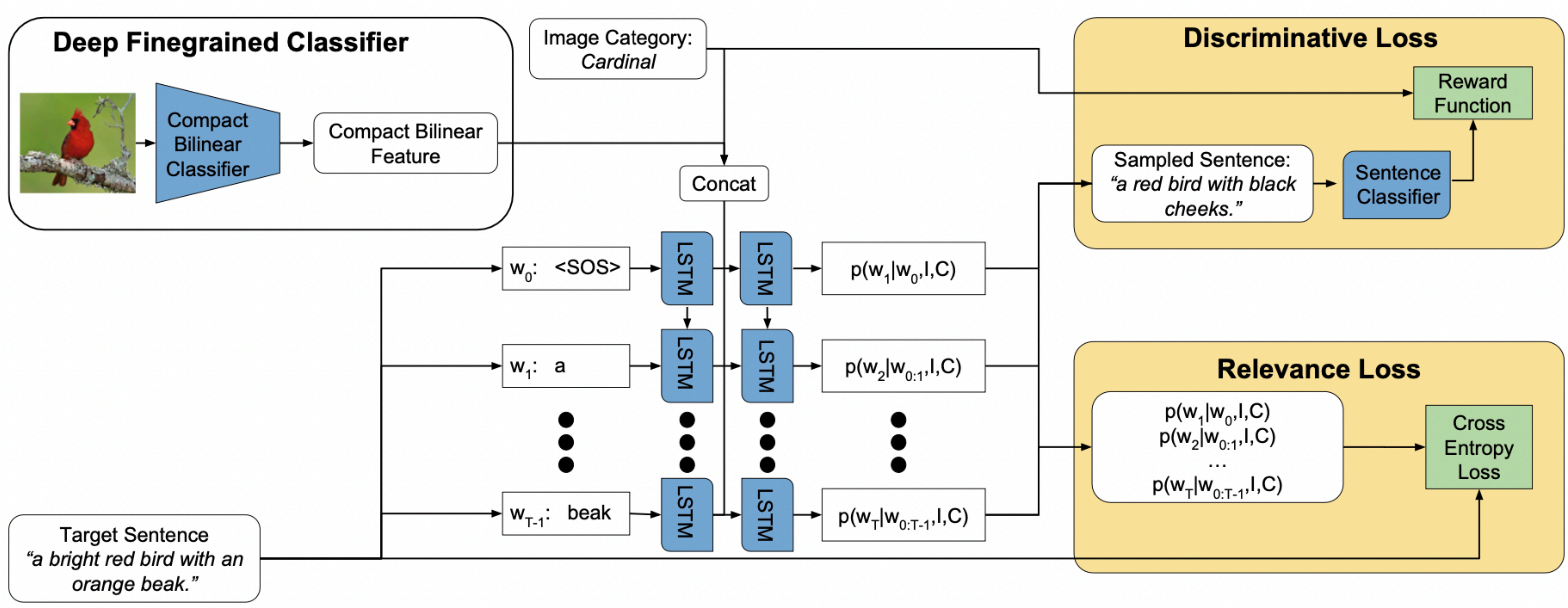




Hendricks et al. (2016)

Generating visual explanations

Model architecture: overview



Hendricks et al. (2016)

Generating visual explanations

Model architecture

- **literal listener:** pretrained LSTM classifier: $P_{L_0}(C \mid c)$
- **literal speaker:** pretrained NIC: $P_{S_0}(c \mid i)$
  - used to produce class labels to condition pragmatic speaker on
  - input for class C to S_1 is average of embeddings for all i belonging to C, produced by literal speaker
- **pragmatic speaker:** trained speaker module $P_{S_1}(c \mid i, C)$
  - trained to maximize objective function:

$$\underbrace{\log P(c \mid i, C)}_{S_0\text{-like caption}} + \underbrace{\log P_{L_0}(C \mid c)}_{\text{information for } L_0 \text{ about category}}$$

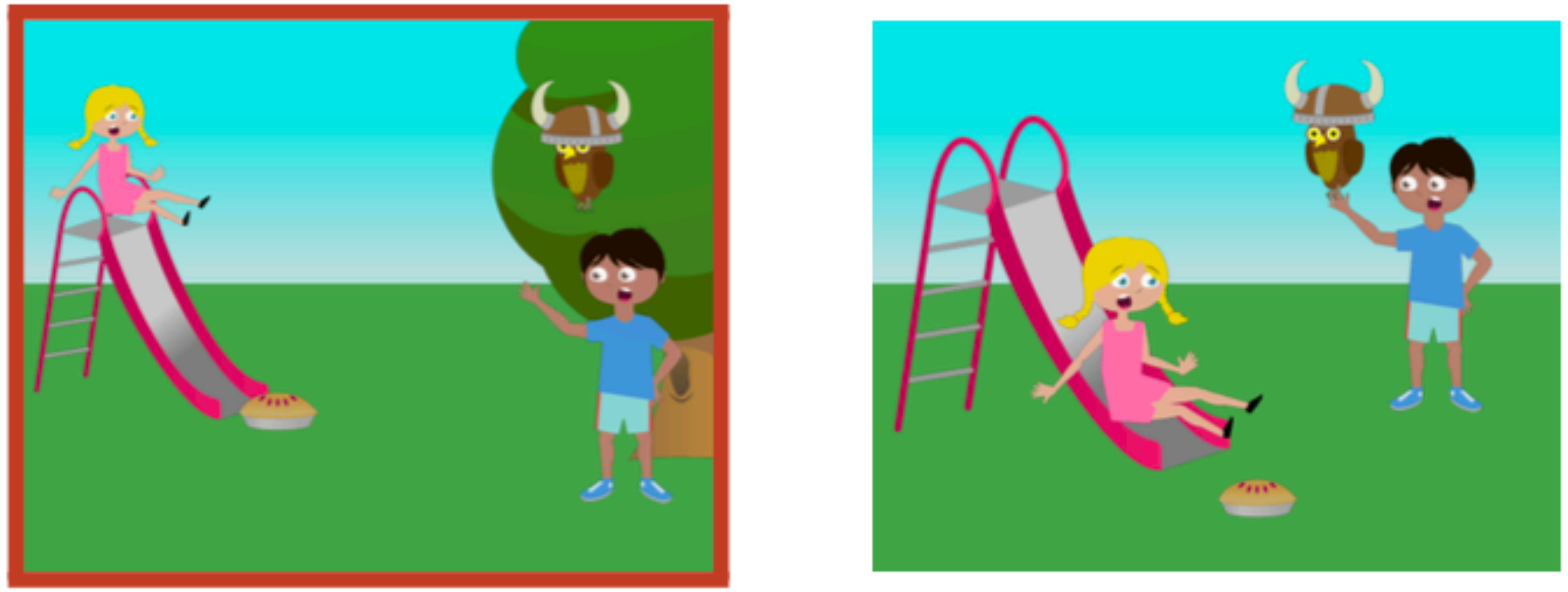
Hendricks et al. (2016)

Reasoning about pragmatics w/ neural listeners and speakers

Andreas & Klein (2016), EMNLP

Neural-Pragmatic Natural Language Generation
for contrastive image captioning

- **goal:** produce caption c that picks out target image i_t over distractor i_d
- **data:** image-caption pairs (i_t, c)
- **literal listener:** pre-trained to maximize
  $P_{L_0}(i_t \mid i_t, i_d, c)$ for all pairs (i_t, c)
- **literal speaker:** pre-trained to maximize
  $P_{S_0}(c \mid i_t)$ for all pairs (i_t, c)
- **pragmatic speaker (reranker):**
  - sample candidates:
    $c_1, \ldots, c_n \sim P_{S_0}(\cdot \mid i_t)$
  - score candidates:
    $s_k = P_{L_0}(i_t \mid i_t, i_d, c_k)^{1-\lambda}\, P_{S_0}(c \mid i_t)^{\lambda}$
  - select caption w/ max. score



(a) target (b) distractor

the owl is sitting in the tree

Andreas & Klein (2016)

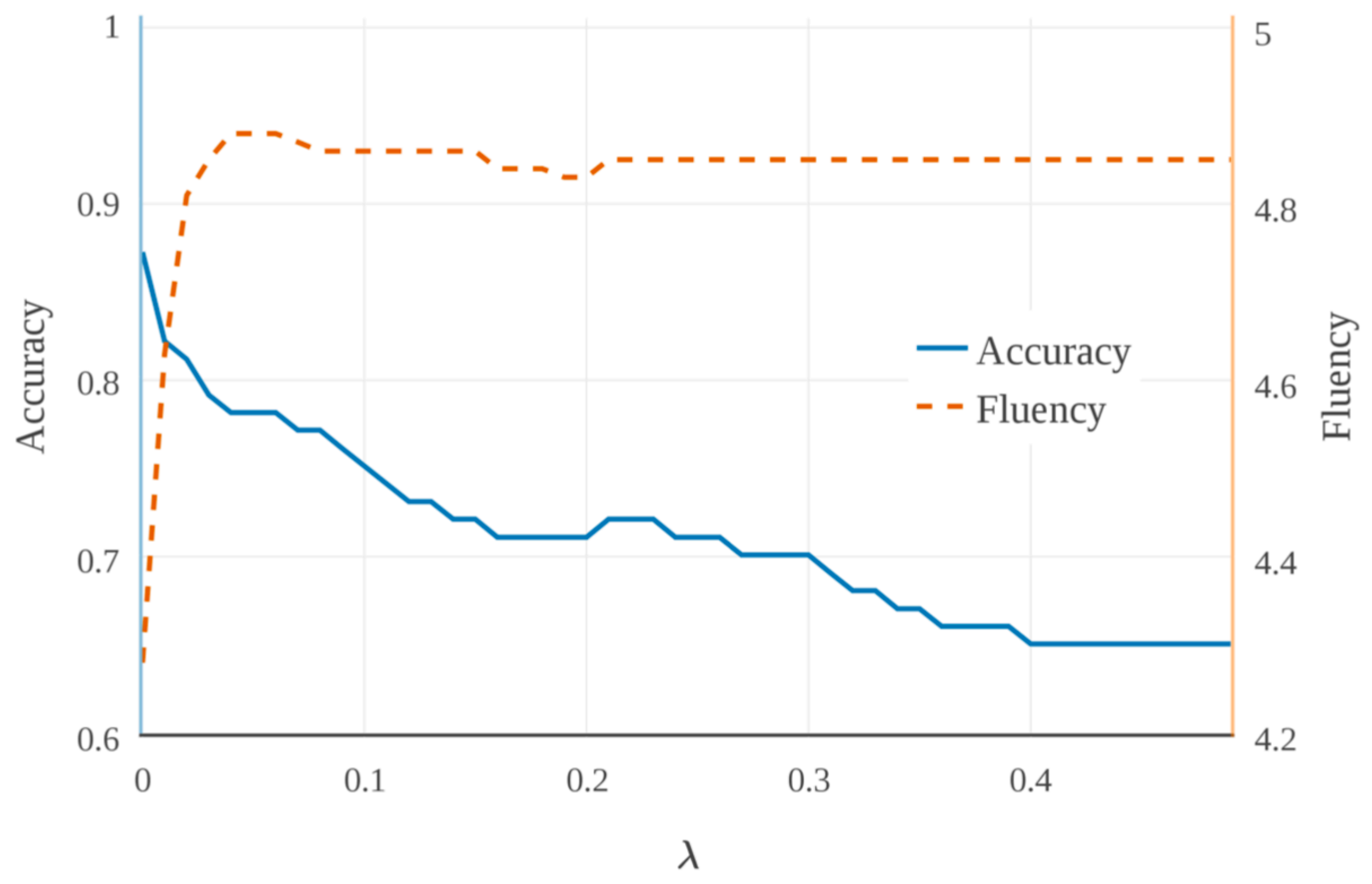
Neural-Pragmatic Natural Language Generation

results

- the more samples we take to score, the higher the accuracy

# samples	1	10	100	1000
Accuracy (%)	66	75	83	85

- accuracy deteriorates with increasing λ



- pragmatic speaker models beats literal speaker baseline, and a reimplementation of the Mao et al. (2015) model

	Dev acc. (%)		Test acc. (%)	
Model	All	Hard	All	Hard
Literal (S0)	66	54	64	53
Contrastive	71	54	69	58
Reasoning (S1)	**83**	**73**	**81**	**68**

Andreas & Klein (2016)

Pragmatically Informative Image Captioning with Character-Level Inference

Cohn-Gordon, Goodman & Potts (2018), NAACL

Incremental neural RSA

model architecture

- **goal:** produce caption c that singles out the target image i_t given a distractor set
- **data:** image-caption pairs (i_t, c)
- **literal speaker:** pre-trained NIC

  $P_{S_0}(w_{1:n} \mid i)$ [neural network]
- **L1-listener:** Bayes rule w/ partial captions

  $P_{L_1}(i \mid w_{1:n}) \propto P_{S_0}(w_{1:n} \mid i)$ [uniform priors]
- **pragmatic speaker (incremental RSA):**

  $P_{S_2}(w_{n+1} \mid i, w_{1:n}) \propto P_{L_1}(i \mid w_{1:(n+1)})^{\alpha} \;\; P_{S_0}(w_{1:(n+1)} \mid i)$
- **granularity:**
  - word-level: each w_n is a full word
  - character-level: each w_n is a single character

S_0 caption: a double decker bus
S_2 caption: a red double decker bus

Cohn-Gordon, Goodman & Potts (2018)

Excursion

formal details of incremental RSA

$$P_{L_1}(i \mid w_{1:n}) = \frac{P(i) \quad P_{S_0}(w_{1:n} \mid i)}{\sum_j P(j) \quad P_{S_0}(w_{1:n} \mid j)}$$

[our reformulation w/ prior]

$$= \frac{P(i) \quad P_{S_0}(w_{1:(n-1)} \mid i) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j P(j) \quad P_{S_0}(w_{1:(n-1)} \mid j) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$

[chain rule]

$$= \frac{\frac{1}{C} P(i) \quad P_{S_0}(w_{1:(n-1)} \mid i) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j \frac{1}{C} P(j) \quad P_{S_0}(w_{1:(n-1)} \mid j) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$

[introducing constant]

$$= \frac{\frac{P(i) \quad P_{S_0}(w_{1:(n-1)} \mid i)}{\sum_k P(k) \quad P_{S_0}(w_{1:(n-1)} \mid k)} \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j \frac{P(j) \quad P_{S_0}(w_{1:(n-1)} \mid j)}{\sum_k P(k) \quad P_{S_0}(w_{1:(n-1)} \mid k)} \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$

[set k to normalization term]

$$= \frac{P(i \mid w_{1:(n-1)}) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j P(j \mid w_{1:(n-1)}) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$

[formulation from the paper]

Excursion

formal details of incremental RSA

$$P_{S_2}(w_{n+1} \mid i, w_{1:n}) \propto \exp\left(\alpha \left(\log P_{L_1}(i \mid w_{1:(n+1)}) - \text{Cost}(w_{1:(n+1)}, i)\right)\right)$$ [vanilla RSA]

$$\propto P_{L_1}(i \mid w_{1:(n+1)})^{\alpha} \exp\left(-\text{Cost}(w_{1:(n+1)}, i)\right)$$ [rules of exponential function]

$$= P_{L_1}(i \mid w_{1:(n+1)})^{\alpha} \; P_{S_0}(w_{1:(n+1)} \mid i)$$ [defining costs via S_0 production]

$$\text{Cost}(w_{1:n}, i) = \log P_{S_0}(w_{1:n} \mid i)^{-\alpha}$$

Upshot:

incremental RSA is, by definition, just plan vanilla RSA
(with a special interpretation of the cost term)

Incremental neural RSA

results

- compare literal and pragmatic models, for character- and word-level incremental predictions
  - but table shows possibly misleading contrast
  - Char S_2 uses beam search for decoding (beam size 10) but Word S_2 uses greedy decoding
  - with greedy decoding Char S_2 scores 61.2% on TS1
  - the advantage could solely come from different decoding

Model	**TS1**	**TS2**
Char S_0	48.9	47.5
Char S_1	**68.0**	**65.9**
Word S_0	57.6	53.4
Word S_1	60.6	57.6

Cohn-Gordon, Goodman & Potts (2018)

Context-aware Captions from Context-agnostic Supervision

Vedantam et al. (2017), CVPR

Emitter-Suppressor model

Task-neutral pre-trained NICs for justification & discriminative captioning

- tasks:
  - **justification:** describe picture by contrasting it against a competitor *class*
  - **discrimination:** describe picture by contrasting it against a competitor *image*
- approach:
  - task-neutral pre-trained NIC
  - novel "**pragmatic beam search**"
  - emitter-suppressor objective function
    - similar but not equivalent to an RSA S_2 model
- data sets:
  - CUB-Justify (novel)
    - extension of the CUB data set w/ new contrastive captions
    - participants described an image in contrast to six images from the contrast class
  - MS-COCO

justification

discrimination

Vedantam et al. (2017)

Emitter-Suppressor model

model architecture

- baseline models (S_0):
  - justification:

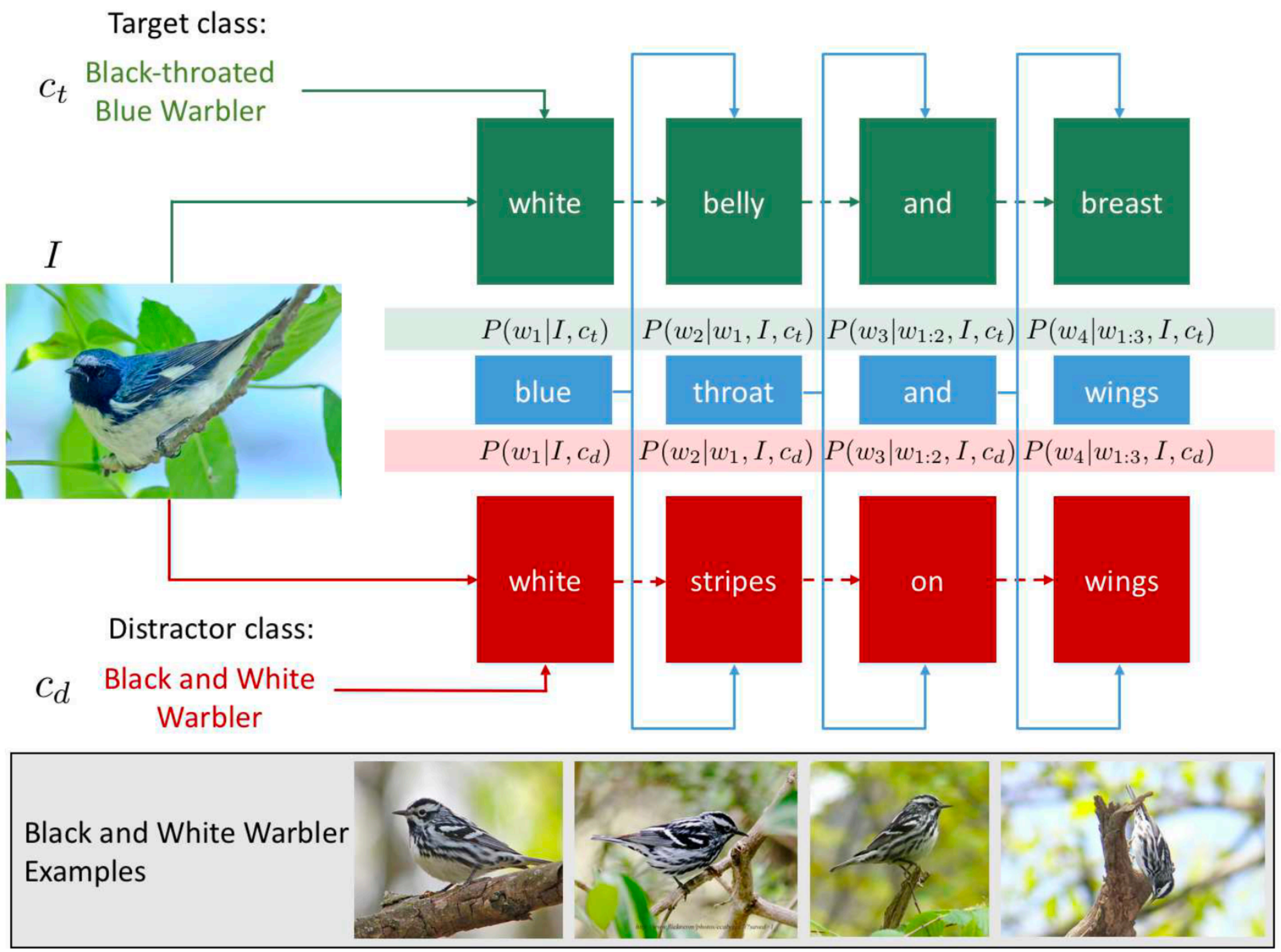
    $P_{S_0}(w_{1:n} \mid i, C_t)$ [caption given image and target class]
  - discrimination:

    $P_{S_0}(w_{1:n} \mid i)$ [caption given image]
- pragmatic speaker ("S_2") (here only for justification):

$$P_{S_2}(w_{1:n} \mid i, C_t, C_d) \propto \lambda \ \log P_{S_0}(w_{1:n} \mid i, C_t) \ + \ (1-\lambda) \ \log \frac{P_{S_0}(w_{1:n} \mid i, C_t)}{P_{S_0}(w_{1:n} \mid i, C_d)}$$

- beam-search maximization:
  - score each proposed word w_{n+1} by **ES objective**:

$$\log \frac{P_{S_0}(w_{1:n} \mid i, C_t)}{P_{S_0}(w_{1:n} \mid i, C_d)^{(1-\lambda)}}$$



Vedantam et al. (2017)

Emitter-Suppressor model
relation to RSA

- the ES-model is formulated only for maximization, but we can define a probabilistic speaker similar to RSA like so:

$$P_{ES}(w_{1:n} \mid i, C) = \text{SM}_\alpha \left(\log \frac{P_{S_0}(w_{1:n} \mid i, C_t)}{P_{S_0}(w_{1:n} \mid i, C_d)^{(1-\lambda)}} \right)$$

- formal results:
  - this model and a vanilla S_2 RSA speaker predict the same ordering on captions if $\alpha = 1$ & $\lambda = 1$
  - predictions are still not identical for $\alpha = 1$ & $\lambda = 1$
  - for other parameter settings, they are not even order equivalent (i.e., could have different arg-max values)
- desideratum / open question:
  - systematically investigate model differences
  - empirically test w/ human subjects

Vedantam et al. (2017)

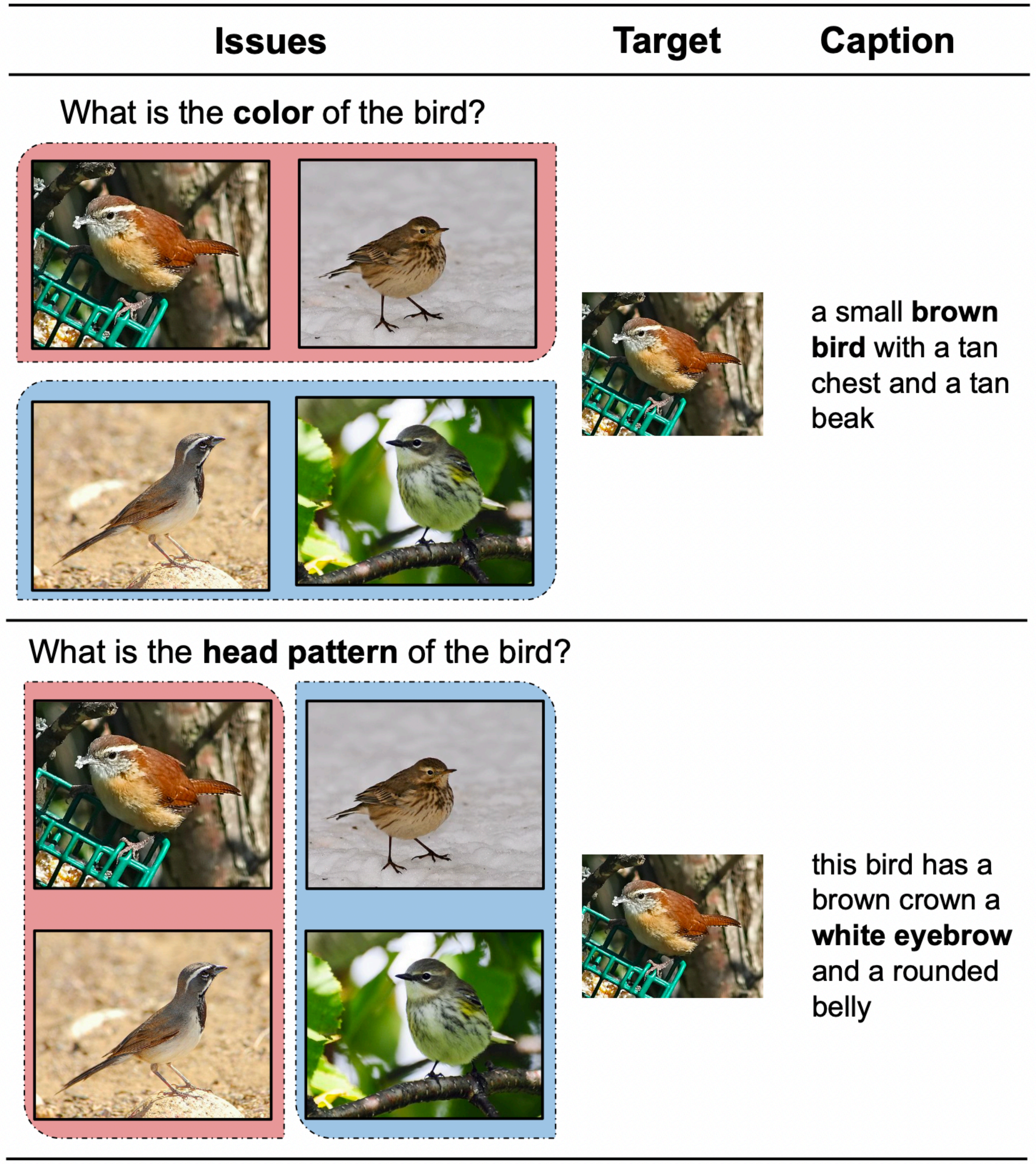
Pragmatic Issue-Sensitive Image Captioning

Nie et al. (2020), EMNLP

Pragmatic Issue-Sensitive Image Captioning

goal and approach

- **goal:** image captions that address a topic question
  - topic question is given by a set of images
- **set-up:** S_0-L_1-S_2 architecture with (pragmatic) beam search, but additional utility components in S_2
  - S_0 is from Hendricks et al. (2016)
- **data:** CUB-captions (Reed et al. 2016)
- additionally: visual question-answering on MS-COCO



Nie et al. (2020)

Pragmatic Issue-Sensitive Image Captioning

model

- **data:** image-caption pairs (i_t, c)
- **issue:** an issue C is a partition of a subset of images
  - $C(i)$ is the element of C that contains i
- **literal speaker:** $P_{S_0}(c \mid i)$ pre-trained NIC [from Hendricks et al. (2016)]
- **L1-listener:** Bayes rule $P_{L_1}(i \mid c) \propto P_{S_0}(c \mid i)$ [uniform priors]
- **pragmatic speakers:**

$$P^X_{S_2}(c \mid i, C) = \text{SM}\left(U^X(i, c, C) + \log P_{S_0}(c \mid i)\right)$$

- **utility functions:** for $X \in \{\emptyset, C, C+H\}$

$$U(i, c, C) = \log P_{L_1}(i \mid c)$$

$$U^C(i, c, C) = \log P_{L_1}\left(C(i) \mid c\right)$$

$$U^{C+H}(i, c, C) = \beta U^C(i, c, C) + (1-\beta)\mathscr{H}\left(P_{L_1}\left(\cdot \mid C(i), c\right)\right)$$

Nie et al. (2020)

Pragmatic Issue-Sensitive Image Captioning

evaluation & results

- automatic assessment of pragmatic adequacy
- human evaluation:
  - 105 participants from MTurk; 13 trials each
  - trials consisted of 110 images and model generations for these

Question: **What is the beak shape?**

Caption: **this is a white bird with black feet and a pointy downward beak**

Select the answer conveyed by the caption, or indicate that the caption doesn't provide an answer:

- ○ **curved_(up_or_down)**
- ○ **dagger**
- ○ **hooked**
- ○ **needle**
- ○ **hooked_seabird**
- ○ **spatulate**
- ○ **all-purpose**
- ○ **cone**
- ○ **specialized**
- ○ **The caption answers the question, but not with one of the above options**
- ○ **The caption does not contain an answer to the question**

no irrelevant features?

	Precision	Recall	F_1
S_0	10.5	21.1	15.5
S_0 Avg	12.1	29.0	17.0
S_1	11.2	21.7	14.8
$S_1^{\mathbf{C}}$	**18.7**	42.5	**25.9**
$S_1^{\mathbf{C}+H}$	16.6	**46.6**	24.5

% or humans considering the issue resolved

Caption Source	Percentage	Size
S_0	20.9	273
S_1	24.5	273
$S_1^{\mathbf{C}}$	42.1	273
$S_1^{\mathbf{C}+H}$	**44.0**	273
Human	33.3	273

training data

Nie et al. (2020)

Multi-agent Communication meets Natural Language

Lazaridou, Potapenko & Tieleman (2020), ACL

Fine-tuning from self-play

Multi-Agent Communication meets Natural Language

- **goal:** task-specific fine-tuning via self-play in multi-agent communication games
- **set-up:**
  - speaker: pre-trained NIC $P_{S_0}(c \mid i)$
  - listener: pretrained image picker: $P_{L_0}(i_t, i_d \mid c)$
  - self-play reference game:
    - speaker and listener repeatedly play reference game
    - update behavioral policies based on success/failure in each round
- different architectures for self-play & update
  - functional or structural learning only
  - both functional & structural learning:
    - fine-tuning via reinforcement learning of S_0 and/or R_0
    - RL-based policy learning for scoring samples from S_0
- problem: **language drift**
  - evolving language is “intelligible” only to the agents

Target Image

Distractor Image

Structural-only learning	
image captioning (§4.2)	
sample	**jenny** is wearing a hat
greedy	mike is wearing a hat
Structural and functional learning	
Gradients from reward affect base captioning model	
reward finetuning (§4.3.1)	
no KL-term	it is camping **camping** [...] camping
with KL-term	mike is sitting on the tent
multi-task learning (§4.3.2)	
$\lambda_s = 0.1$	mike is jenny on the the tent
$\lambda_s = 1$	mike is sitting on the ground
Reranking (§4.3.3), base captioning model unchanged	
PoE, $\lambda_s = 0$	the tent is in the tree
PoE, $\lambda_s = 1$	mike and jenny are sitting **on the ground**
noisy channel	jenny is wearing a **funny hat**

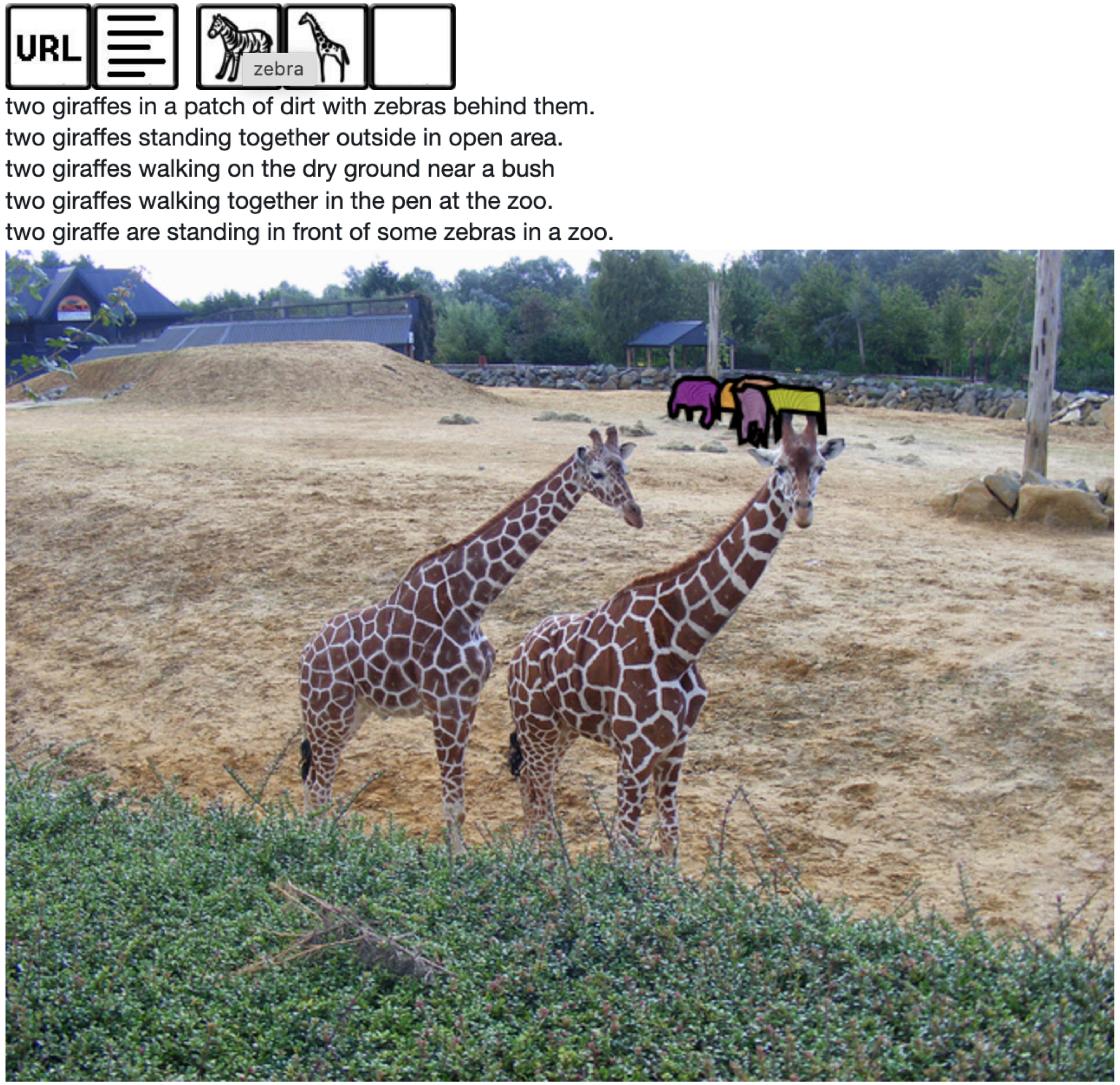
Lazaridou et al. (2020)

Data Sets

MSCOCO

large data set w/ images, captions & labelled-objects

- > 300k images with:
  - captions
  - bounding boxes for 80 objects w/ labels
    - things (concrete objects) and stuff (background elements)
- **URL:** https://cocodataset.org



Yin et al. (2014), “Microsoft COCO: Common Objects in Context”, ECCV

Google Refexp

referential expressions for objects in MS-COCO images

- subset of images from MS-COCO w/ additional referential expressions for objects in the images
- > 26k images with 54k target objects
  - each object types occurs 2-4 times in the picture
  - all objects of that type are sufficiently salient
  - bounding boxes and labels for objects (from MS-COCO)
- ~1.9 referential expressions per target object
  - obtained from MTurk human annotation
    - human producer types referential expression E
    - human interpreter tries to identify target object based on E
    - if successful E is added to data set, if not discarded
- **URL:** Google Refexp

The black and yellow backpack sitting on top of a suitcase.

A yellow and black back pack sitting on top of a blue suitcase.

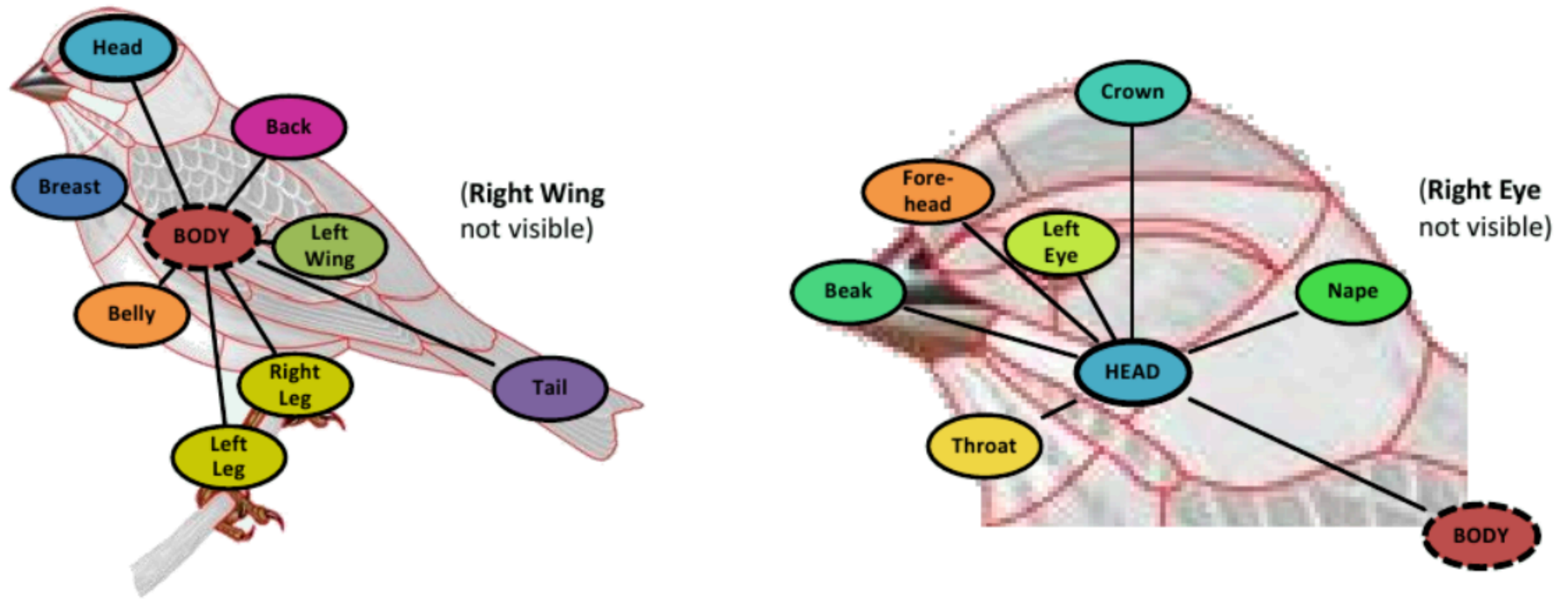
An apple desktop computer.

The white IMac computer that is also turned on.

Mao et al. (2016) "Generation and Comprehension of Unambiguous Object Descriptions" , CVPR

Caltech-UCSD Birds

w/ captions and justifications



- original CUB
  - ~11.8k images of 200 bird species
  - taxonomic information: order, family, genus, species
  - 312 binary attributes (e.g., bill shape)
  - bounding boxes, attributes & part locations
- CUB-captions extension (Reed et al. 2016)
  - five captions per picture
  - human captioners did not have access to attribute info
- CUB-justify extension (Vedantam et al. 2017)
  - obtained from MTurk human annotation
    - human producer types description of a target image from class X in contrast to six images from competitor category Y
- **URLs:** CUB, CUB-caption, CUB-justify

Part	Attributes	Part	Attributes	Part	Attributes
Beak	*HasBillShape, HasBillColor , HasBillLength*	**Back**	*HasBackColor, HasBackPattern*	**Breast**	*HasBreastPattern, HasBreastColor*
Belly	*HasBellyPattern, HasBellyColor*	**Fore-head**	*HasForehead Color*	**Bird** (all parts)	*HasSize, HasShape*
Throat	*HasThroatColor*	**Nape**	*HasNapeColor*	**Head**	*HasHeadPattern*
Crown	*HasCrownColor*	**Eye**	*HasEyeColor*	**Leg**	*HasLegColor*
Tail	*HasUpperTailColor, HasUnderTailColor, HasTailPattern, HasTailShape*	**Wing**	*HasWingPattern, Has WingColor, HasWingShape*	**Body**	*HasUnderpartsColor, HasUpperPartsColor, HasPrimaryColor*

Attribute Annotation
Has_Bill_Shape::All-purpose
Has_Wing_Color::Brown
Has_Wing_Color::Rufous
Has_Back_Color::Brown
Has_Head_Pattern::Eyebrow
Has_Size::Small

Wah et al. (2011) "Caltech-UCSD Birds-200-2011" , technical report

Abstract scenes

- 10k synthetic images w/ ~ 6 captions per image
- generation procedure:
  - **original scenes:** ~1k scenes with 10 descriptions each:
    - based on 80 pieces of clip art
    - first set of human participants instructed to "*create an illustration for a children's story book by creating a realistic scene from the clip art*"
    - second set of participants created one description for each scene
  - **similar scenes:**
    - for each written description humans created 10 scenes (see pic)
  - **additional labels:**
    - human annotators provide ~6 description for each of the resulting 10k scenes
- **URL:** Abstract Scenes

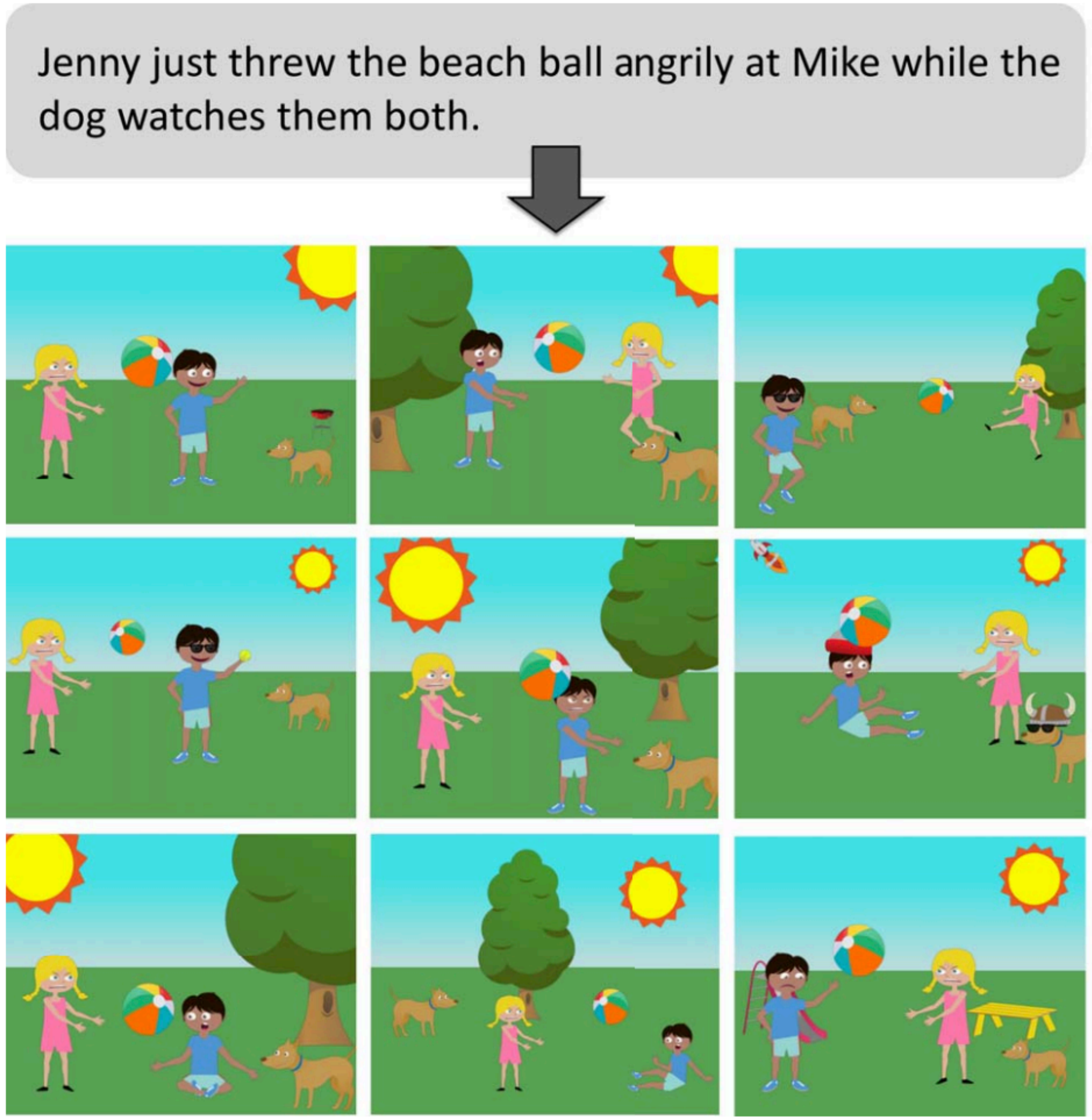


Figure 1. An example set of semantically similar scenes created by human subjects for the same given sentence.

Zitnick et al. (2013) "Bringing Semantics Into Focus Using Visual Attention", CVPR