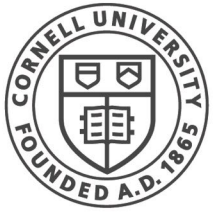
Language and Reasoning Diversity in Grounded Natural Language Understanding

Yoav Artzi

SiVL, NAACL 2019



Cornell CIS
Computer Science

Today

Understanding

NLVR

NLVR2 (*nlvr.ai*)

- Robustness to biases
- Language and reasoning diversity

Acting

Touchdown (*touchdown.ai*)

DRIF

- Real-life input
- Robotic agents

Biases and Reasoning Diversity

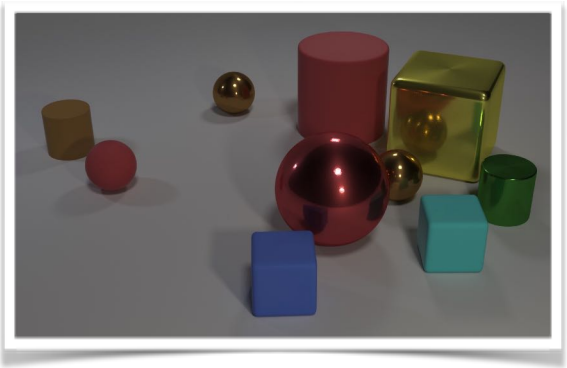
VQA

What is the dog carrying?

Stick

- Implicit biases
- Relatively simple language

CLEVR



Are there an equal number of large things and metal spheres?

Yes

NLVR

there are exactly three squares not touching any edge

TRUE

Natural Language Visual Reasoning (NLVR)

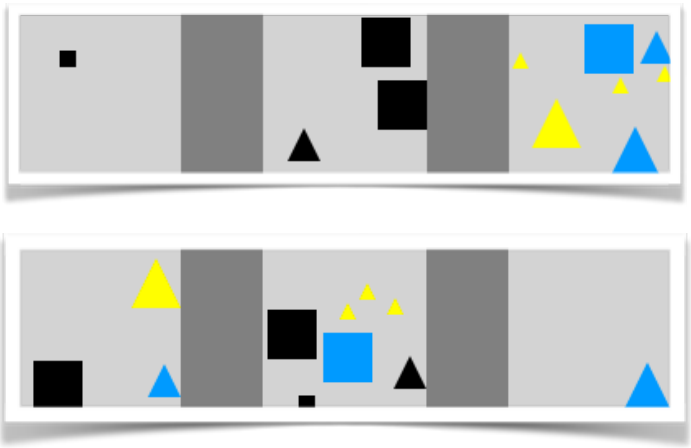
there are exactly three squares not touching any edge

TRUE

- Isolates compositional reasoning problem
- Box structure encourages set and comparison reasoning
- Controlled environment → focus sentences on specific phenomena
- Compare and contrast for balanced data
- But: synthetic vision and limited lexical diversity

Natural Language Visual Reasoning (NLVR)

there are exactly three squares not touching any edge

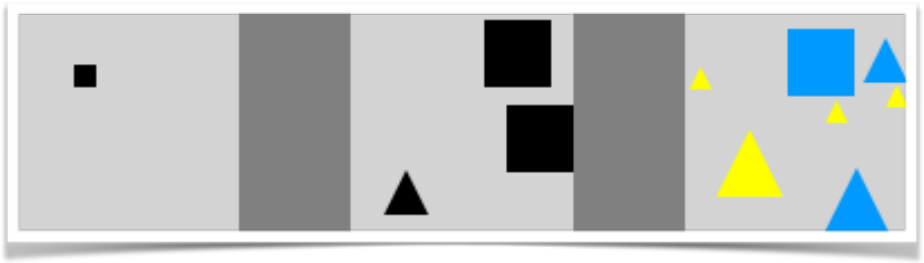


TRUE

FALSE

- Isolates compositional reasoning problem
- Box structure encourages set and comparison reasoning
- Controlled environment → focus sentences on specific phenomena
- Compare and contrast for balanced data
- But: synthetic vision and limited lexical diversity

Natural Language Visual Reasoning (NLVR)



there are exactly three squares not touching any edge

TRUE

How to generalize this type of data to real images?

- No control of image content
- No box structure for set reasoning
- Can't generate images for compare and contrast

Natural Language Visual Reasoning *for Real* (NLVR2)

One image shows exactly two brown acorns in back-to-back caps on green foliage

Task: Determine whether the sentence is true or false about the pair of images

Natural Language Visual Reasoning *for Real* (NLVR2)

One image shows exactly two brown acorns in back-to-back caps on green foliage

FALSE

Task: Determine whether the sentence is true or false about the pair of images

NLVR2

One image shows exactly two brown acorns in back-to-back caps on green foliage

FALSE

- Re-creates the NLVR setup with real web images
- Natural language data
- Paired images analogous to boxes
- Compare and contrast to create balanced data

NLVR2

One image shows exactly two brown acorns in back-to-back caps on green foliage

FALSE

TRUE

FALSE

TRUE

Data Collection

- Collecting images using search engines
- Sentence writing using compare and contrast
- Validation

Image Collection

1. **Pick 124 synsets from ImageNet**
   Chose synsets that would often appear multiple times in one image: e.g., acorn >> sump pump

   - Allows use of ImageNet models and tools
   - Allows for weak annotation of image content

Image Collection

1. Pick 124 synsets from ImageNet
2. **Generate and execute search queries and get similar images**
   Combine synset names with numerical phrases, hypernyms, and similar words

Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. **Remove low-quality images**
   Don't contain synset, drawings, inappropriate content

Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. Remove low-quality images
4. **Construct sets of eight images**
   Each set must contain at least three *interesting* images (e.g., multiple objects)

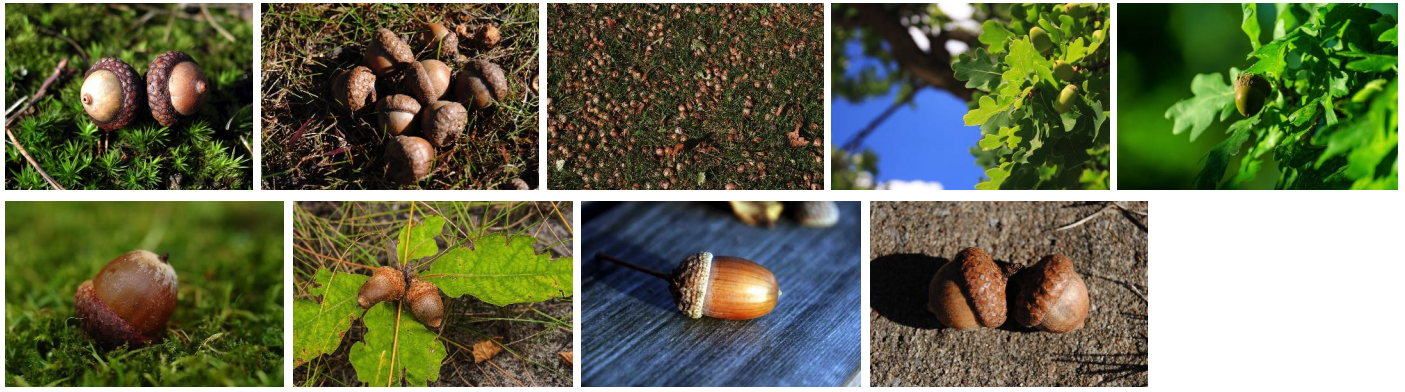


Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. Remove low-quality images
4. **Construct sets of eight images**
   Each set must contain at least three interesting images (e.g., multiple objects)

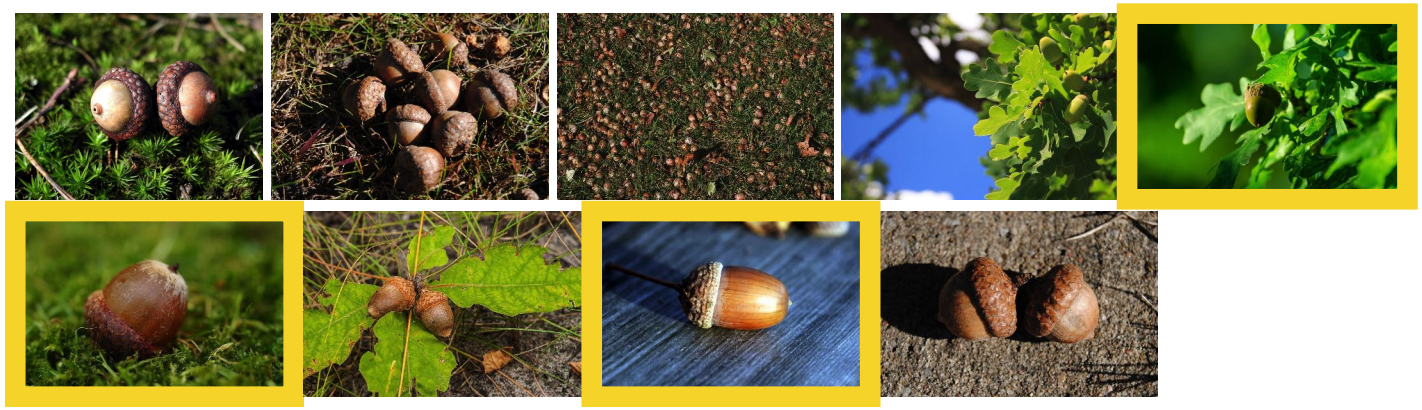


Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. Remove low-quality images
4. **Construct sets of eight images**
   Each set must contain at least three interesting images (e.g., multiple objects)

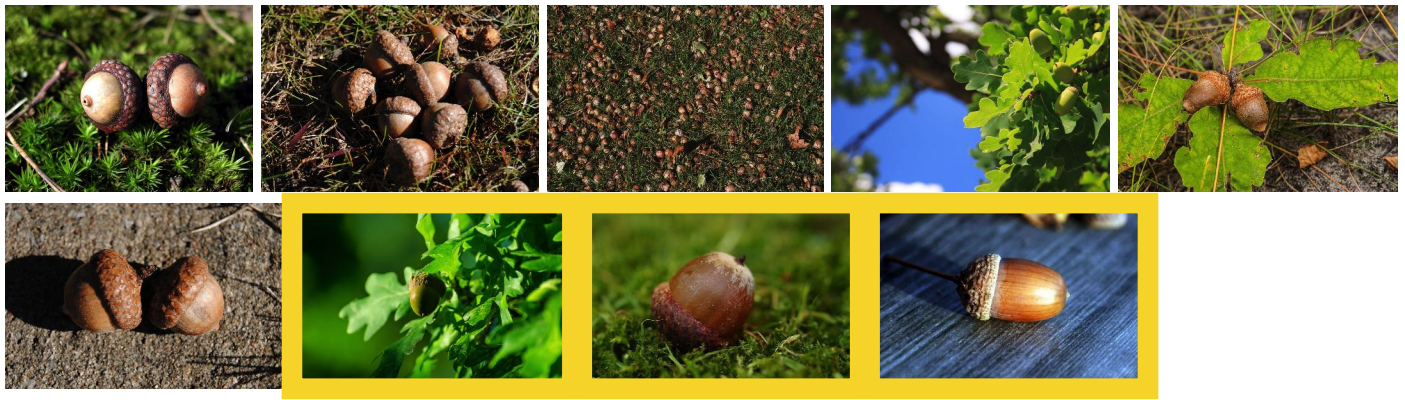


Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. Remove low-quality images
4. **Construct sets of eight images**
   Each set must contain at least three interesting images (e.g., multiple objects)

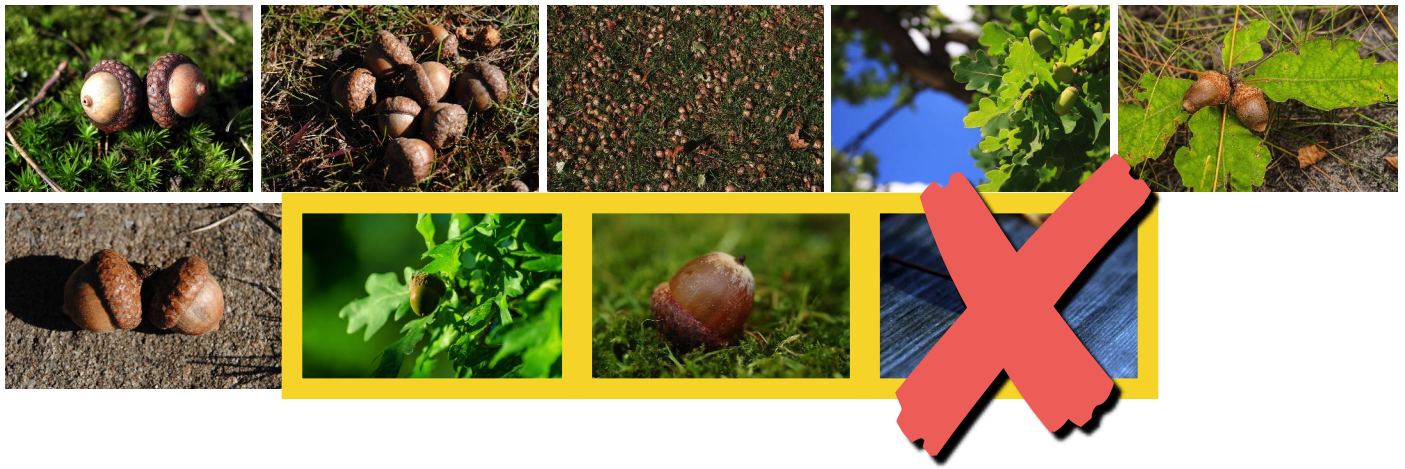
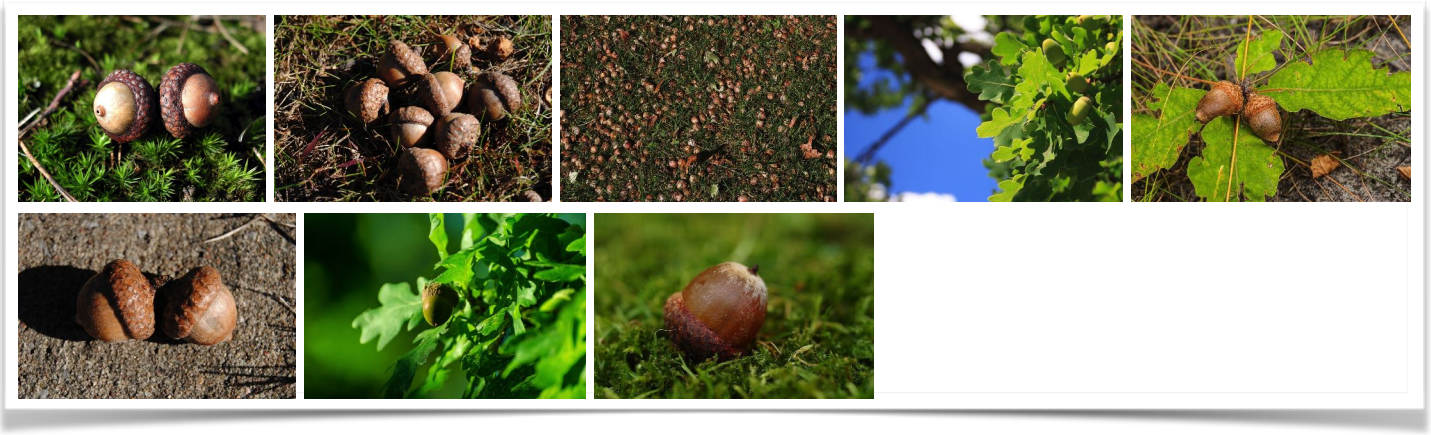


Image Collection

1. Pick 124 synsets from ImageNet
2. Generate and execute search queries and get similar images
3. Remove low-quality images
4. Construct sets of eight images
   Each set must contain at least three interesting images (e.g., multiple objects)



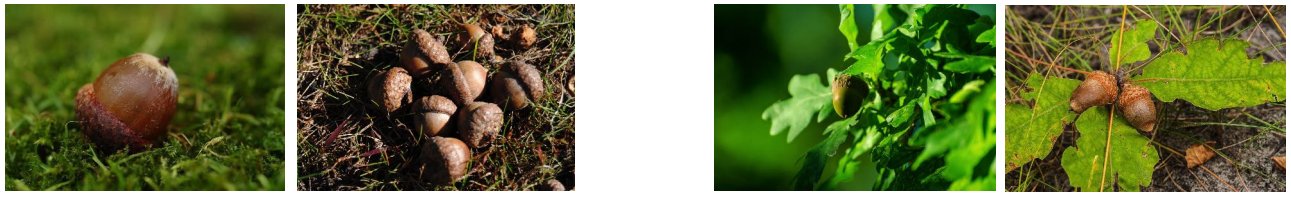
Set of eight images

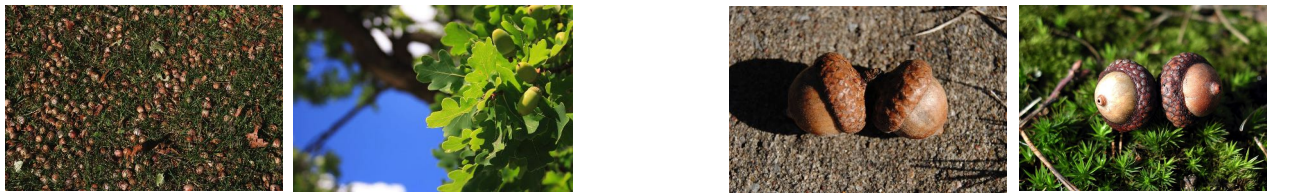
Sentence Writing

5. **Display a set of randomly paired images**

6. Ask workers to select two pairs

7. Workers write a sentence true about the selected pairs, but false about the others



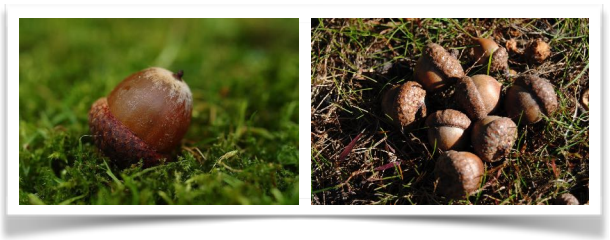


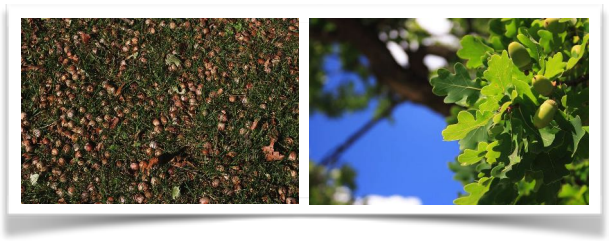
Sentence Writing

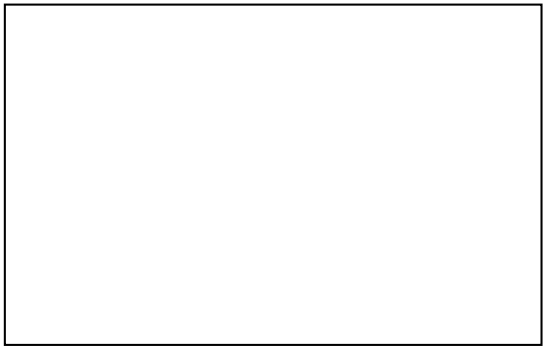
5. **Display a set of randomly paired images**

6. Ask workers to select two pairs

7. Workers write a sentence true about the selected pairs, but false about the others





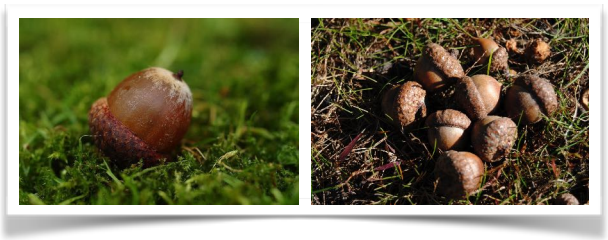


Sentence Writing

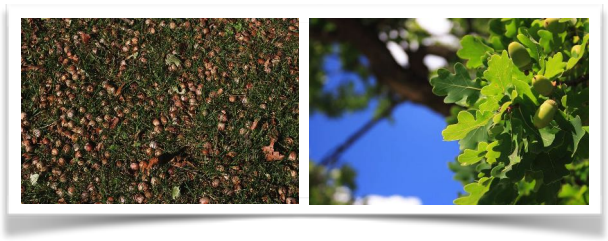
5. Display a set of randomly paired images

6. **Ask workers to select two pairs**

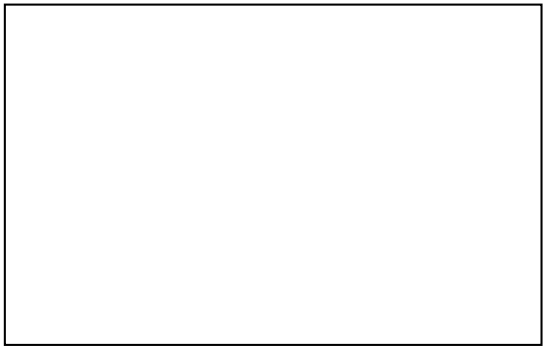
7. Workers write a sentence true about the selected pairs, but false about the others



✔



✔

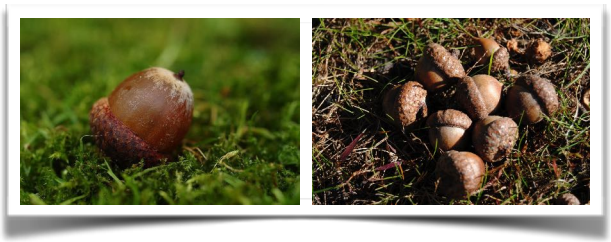


Sentence Writing

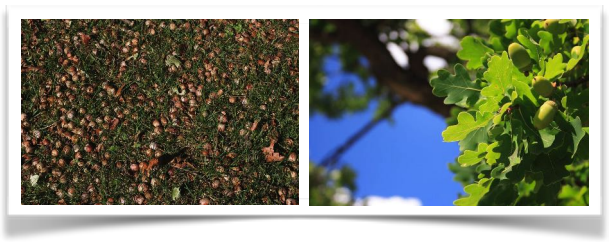
5. Display a set of randomly paired images

6. Ask workers to select two pairs

7. **Workers write a sentence true about the selected pairs, but false about the others**



✔



✔

One image shows exactly two brown acorns in back-to-back caps on green foliage

Validation

8. **Show each images/sentence pair to another worker and ask them to label it**

One image shows exactly two brown acorns in back-to-back caps on green foliage

TRUE

FALSE ✔

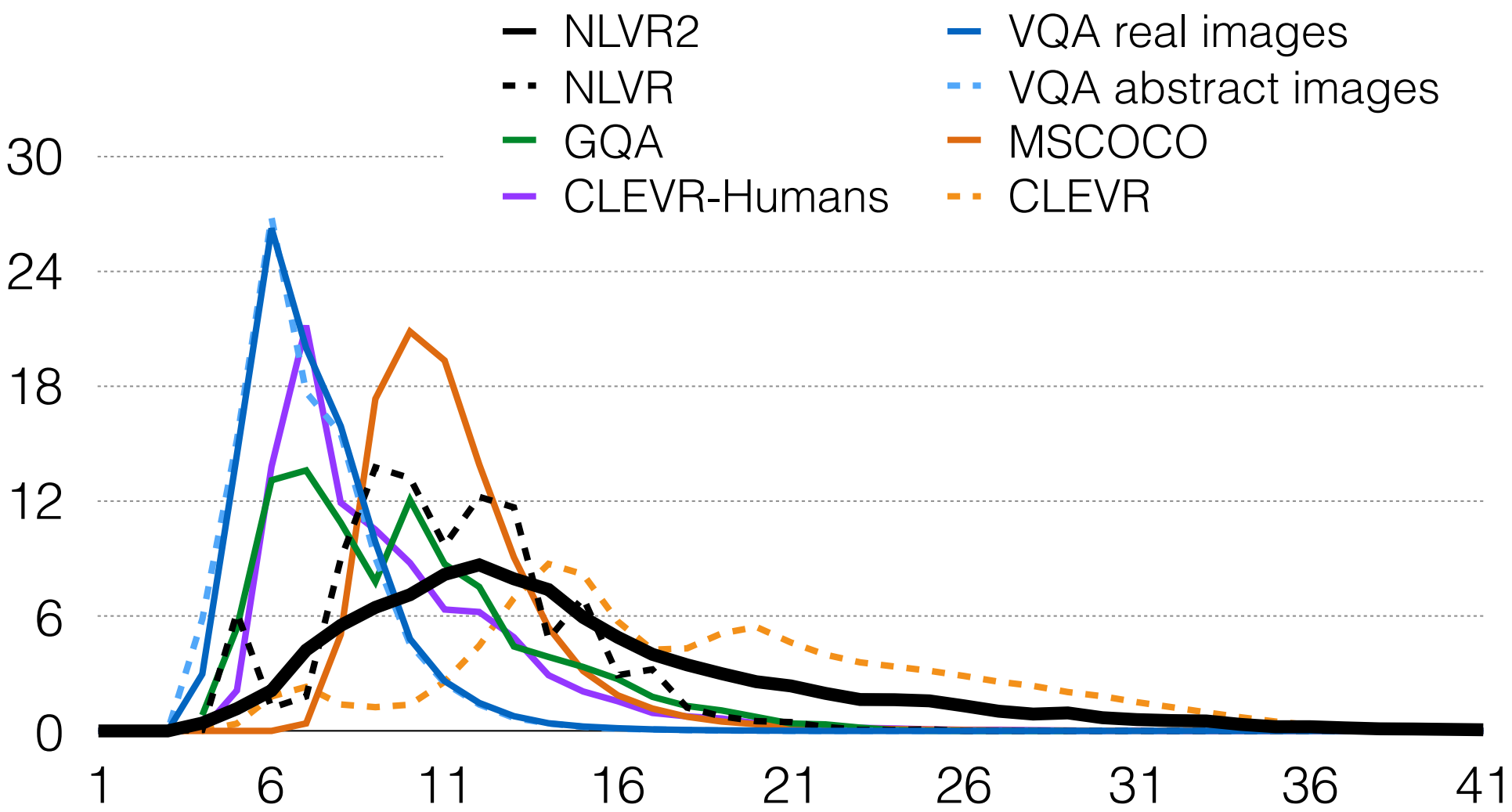
Statistics

- **107,296 total examples**
  - 29,680 unique sentences
  - 127,506 unique images
  - 80% train, 20% evenly split to dev and two test sets
- **Agreement:** near perfect ($\alpha = 0.912$, $\kappa = 0.889$)
- **Total cost:** $19,282.99
- **Average sentence length:** 14.8 tokens
- **Vocabulary size:** ~7,500 word types

Related Resources

	Task	Real Images	Natural Language
VQA	QA	✔	✔
COCO Captions	Caption generation	✔	✔
CLEVR	QA	✗	✗
CLEVR-Humans	QA	✗	✔
GQA	QA	✔	✗
NLVR	Binary classification	✗	✔
NLVR2	Binary classification	✔	✔

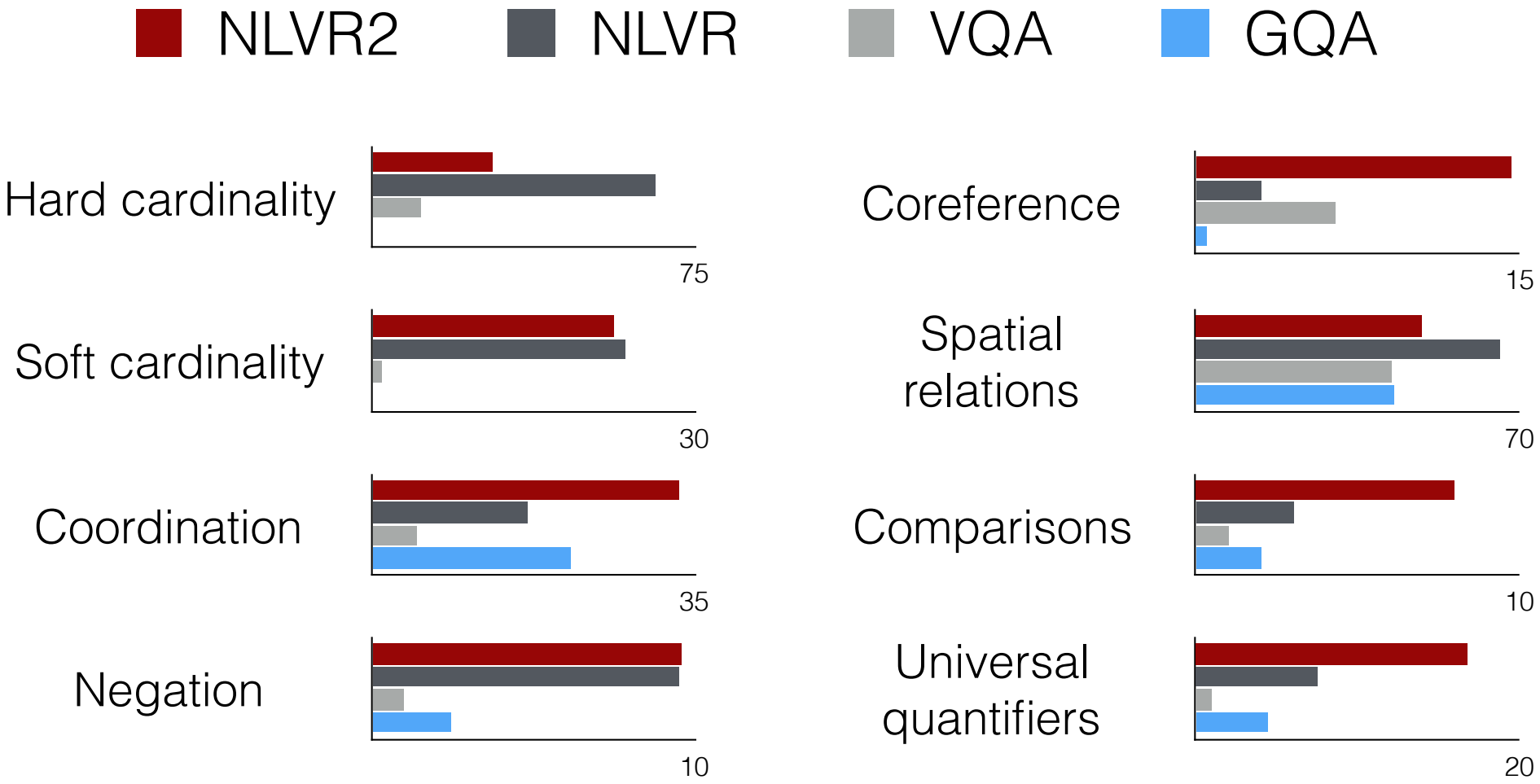
Sentence Length



Linguistic Analysis

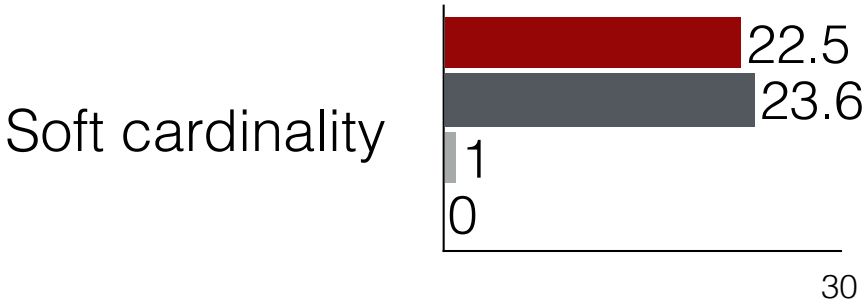
- Analyze 13 semantic and syntactic categories
- Sampled 800 sentences
- Compare to 200 sentences from GQA, VQA, and NLVR
- Release scripts to break down system performance according to categories

Linguistic Analysis



Soft Cardinality

NLVR2 NLVR VQA GQA



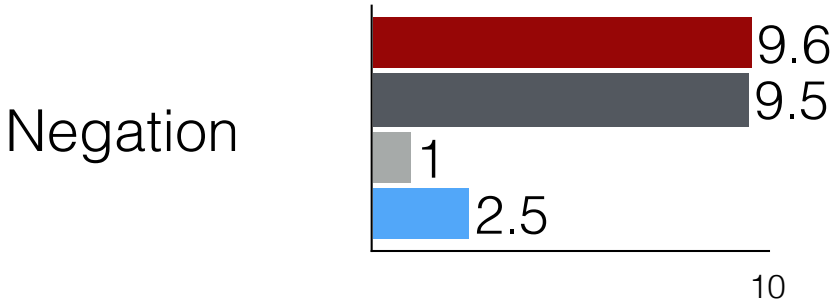
TRUE

One image contains a single vulture in a standing pose with its head and body facing leftward, and the other image contains a group of <u>at least eight</u> vultures.

FALSE

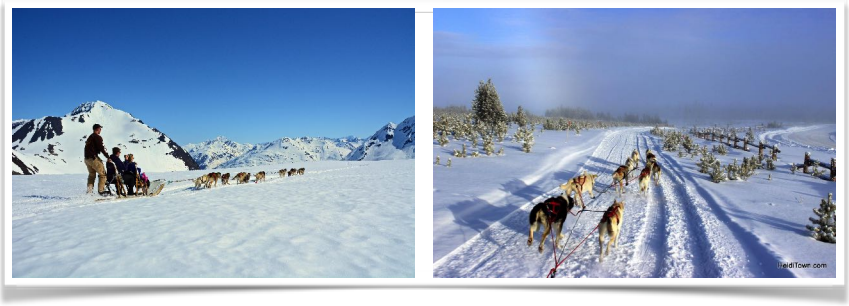
Negation

NLVR2 NLVR VQA GQA



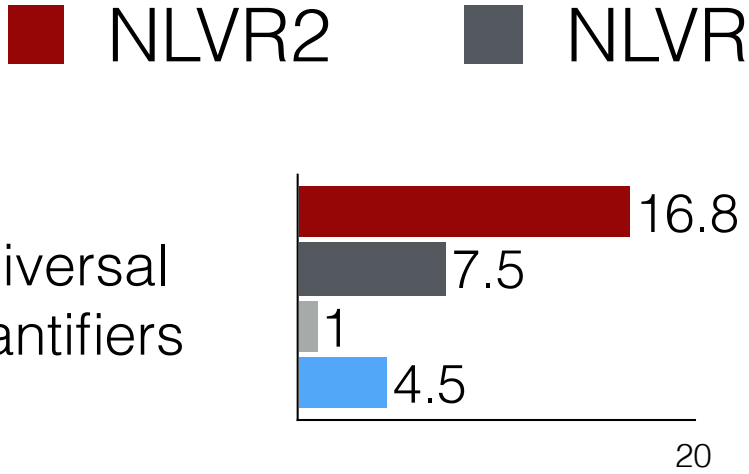
One dog sled team is moving and one is not

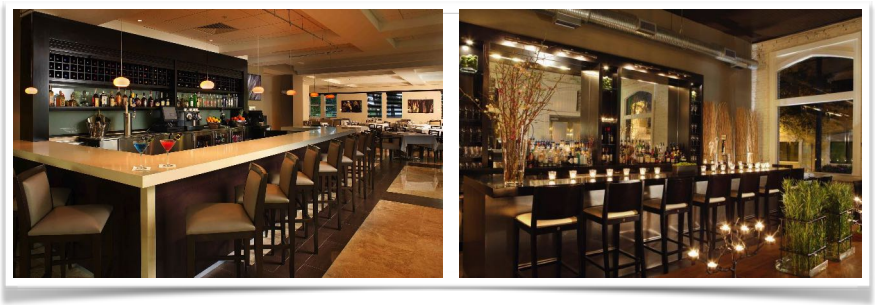
TRUE



FALSE

Universal Quantifiers





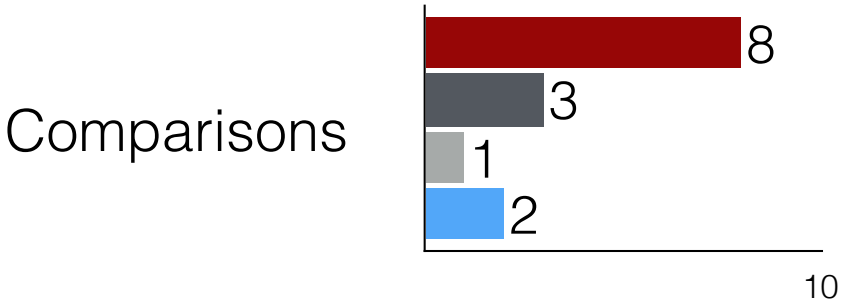
TRUE

<u>All</u> the chairs have backs

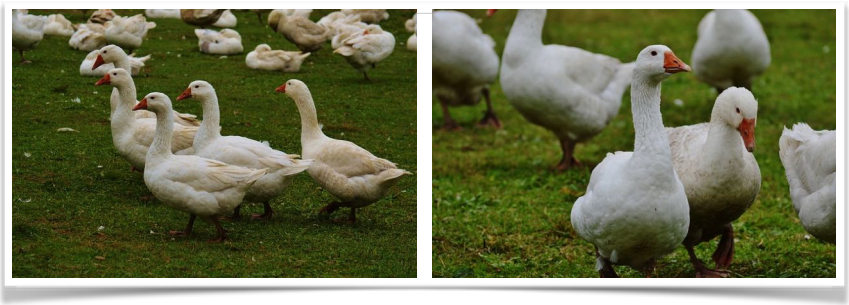
FALSE

Comparisons

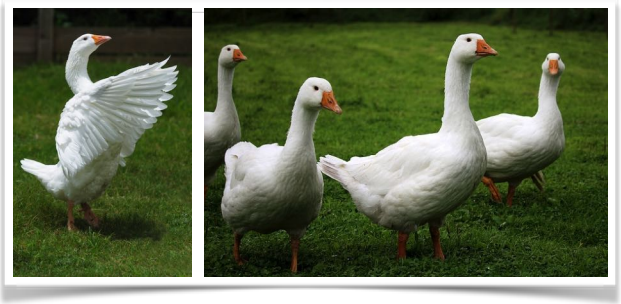
NLVR2 NLVR VQA GQA



There are <u>more</u> birds in the image on the left <u>than</u> in the image on the right



TRUE



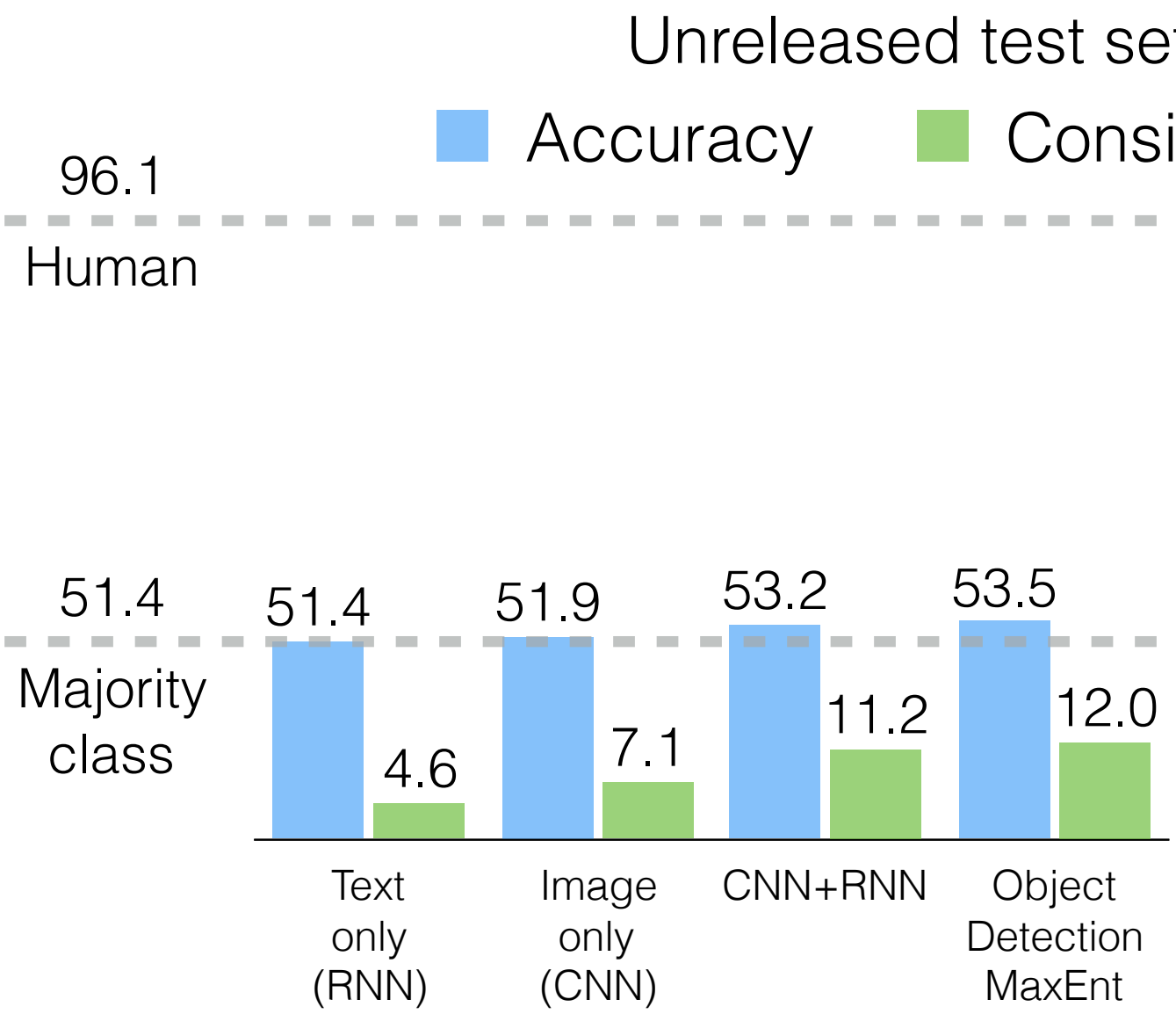
FALSE

Evaluation

- Accuracy
- Consistency
  - Proportion of unique sentences for which predictions are correct for all paired images

[Goldman et al. 2018]

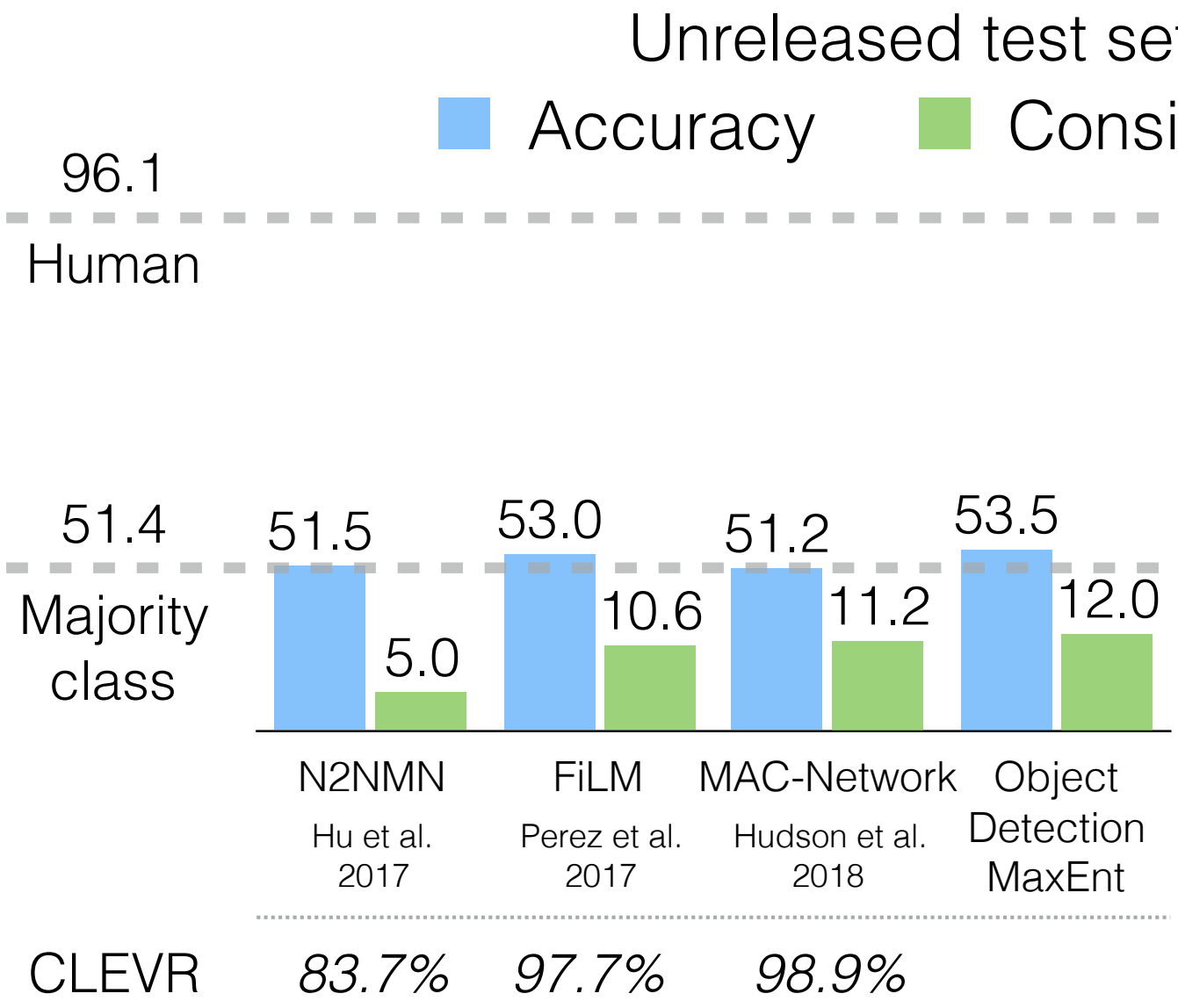
Baselines



- Robust to single-modality biases
- MaxEnt on top of detector does best

Accuracy and consistency are not to scale

SOTA Visual Reasoning



- SOTA methods perform poorly
- CLEVR-NLVR2 performance mismatch

Accuracy and consistency are not to scale

Today

Understanding

NLVR

NLVR2 (*nlvr.ai*)

- Robustness to biases
- Language and reasoning diversity

Acting

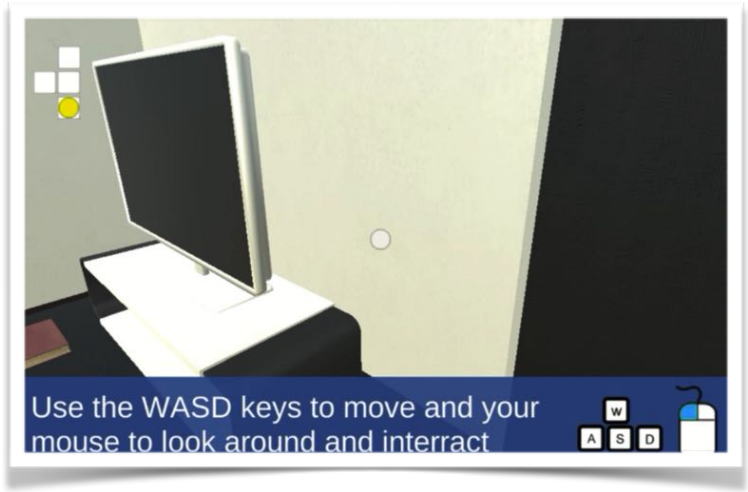
Touchdown (*touchdown.ai*)

DRIF

- Real-life input
- Robotic agents

Realistic Environments

- Most research on instruction following uses simple simulated environments
- Existing physical environments are simple and built in the lab
- Real-life environments are both visually and distributionally different

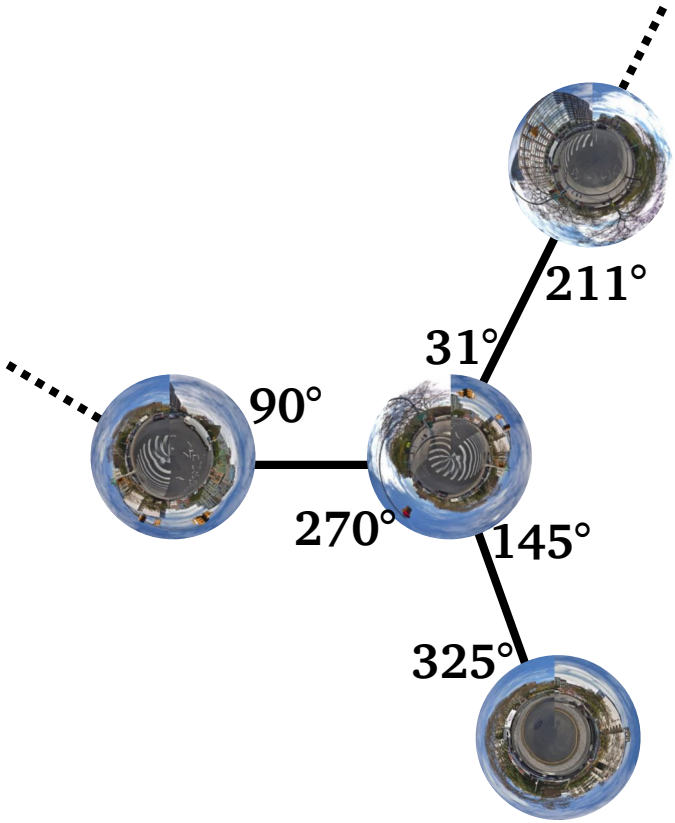


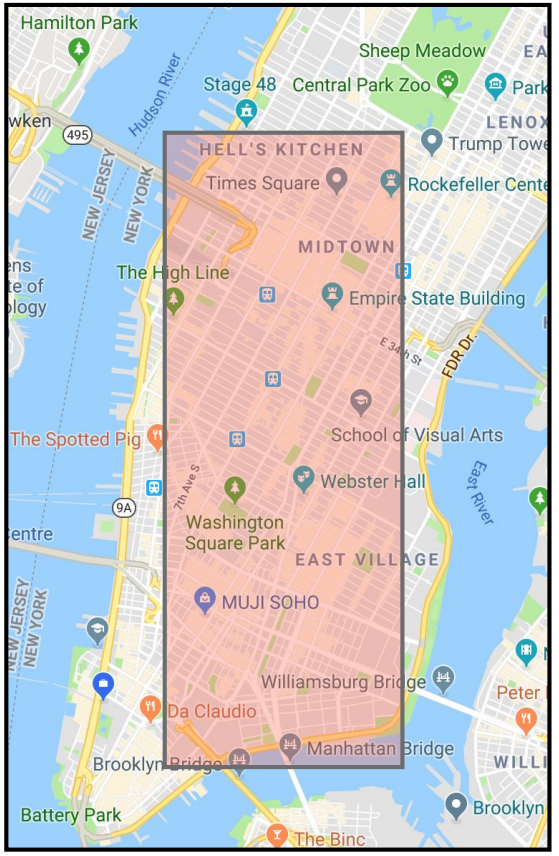
Take watermelon, oranges, and cucumber from the counter. Put them …

Chalet [Yan et al. 2018; Misra et al. 2018]

The Environment

- Google Street View panoramas
- 29,941 panoramas
- 61,319 edges
- 122,638 states for discrete navigation





Task-focused Navigation

- Writing task: instruct to follow a path and describe the location of an object they hide
- The focused task makes the instruction more natural for the writer
- Guide workers not to count intersections and not to use text and store names
- What do we hide?


CORNELL

Example 1

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

Backup

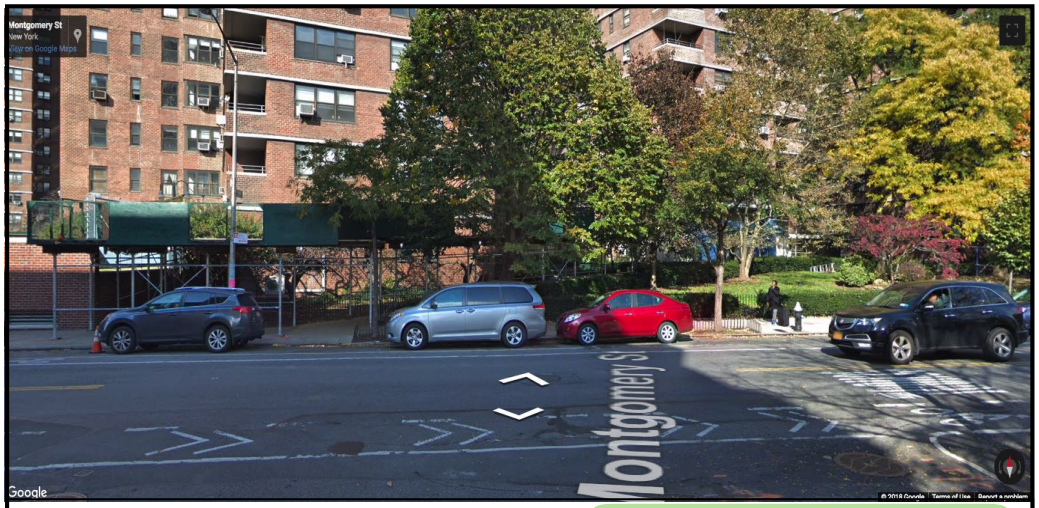
Task-focused Navigation

- This formulation allows for multiple tasks:
  - **Navigation** only: given instruction and a starting point, navigate to the goal position
  - **Spatial description resolution (SDR)** only: given a sentence and a panorama, find Touchdown
  - The complete task: navigate first, and then find Touchdown

Data Collection

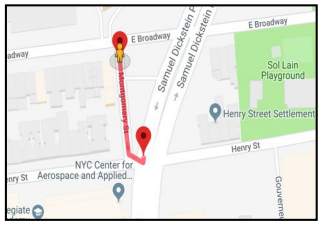
- A sequence of four tasks on Mechanical Turk
  - Writing, propagation, validation, and segmentation
- Workers use a customized StreetView environment

Task I: Writing



Place Touchdown

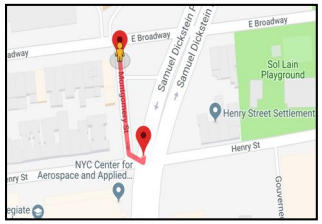
Can’t Place Touchdown



Task I: Writing

Place Touchdown

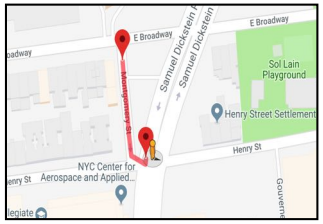
Can't Place Touchdown



Turn so that the trees are to your left. At the first intersection, turn left and stop.

Place Touchdown

Can't Place Touchdown



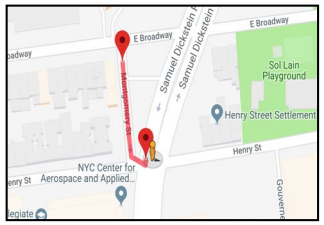
Task I: Writing

e

d

Place Touchdown

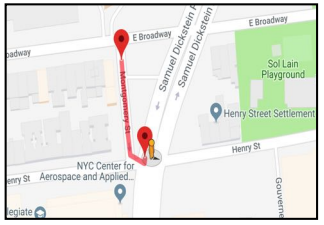
Can't Place Touchdown



Turn so that the trees are to your left. At the first intersection, turn left and stop. **Touchdown is on top of the blue mailbox on the right hand corner.**

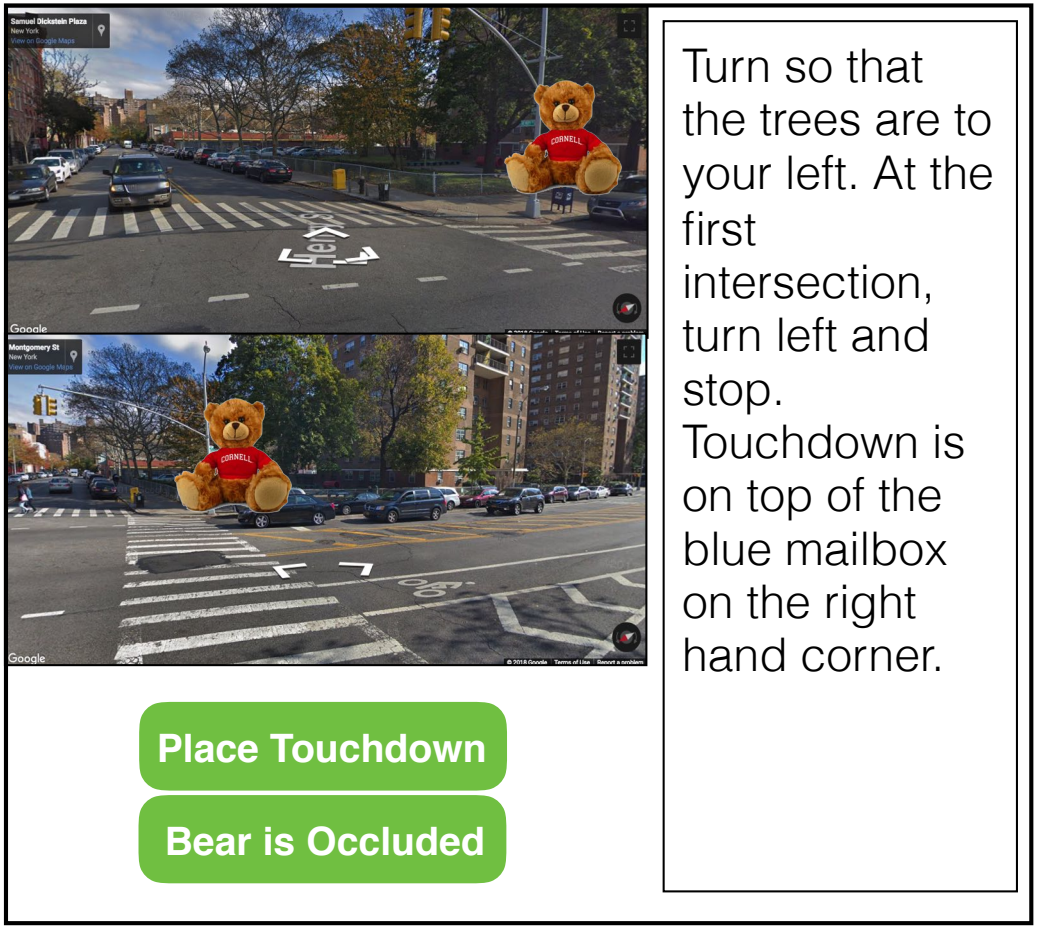
Place Touchdown

Can't Place Touchdown



Task II: Propagation

- Touchdown position may be visible from multiple panoramas
- We propagate the location to neighboring panoramas



Task III: Validation

- Validate instruction by finding Touchdown
- Easy to verify
- Give bonuses to original writer and follower if successful

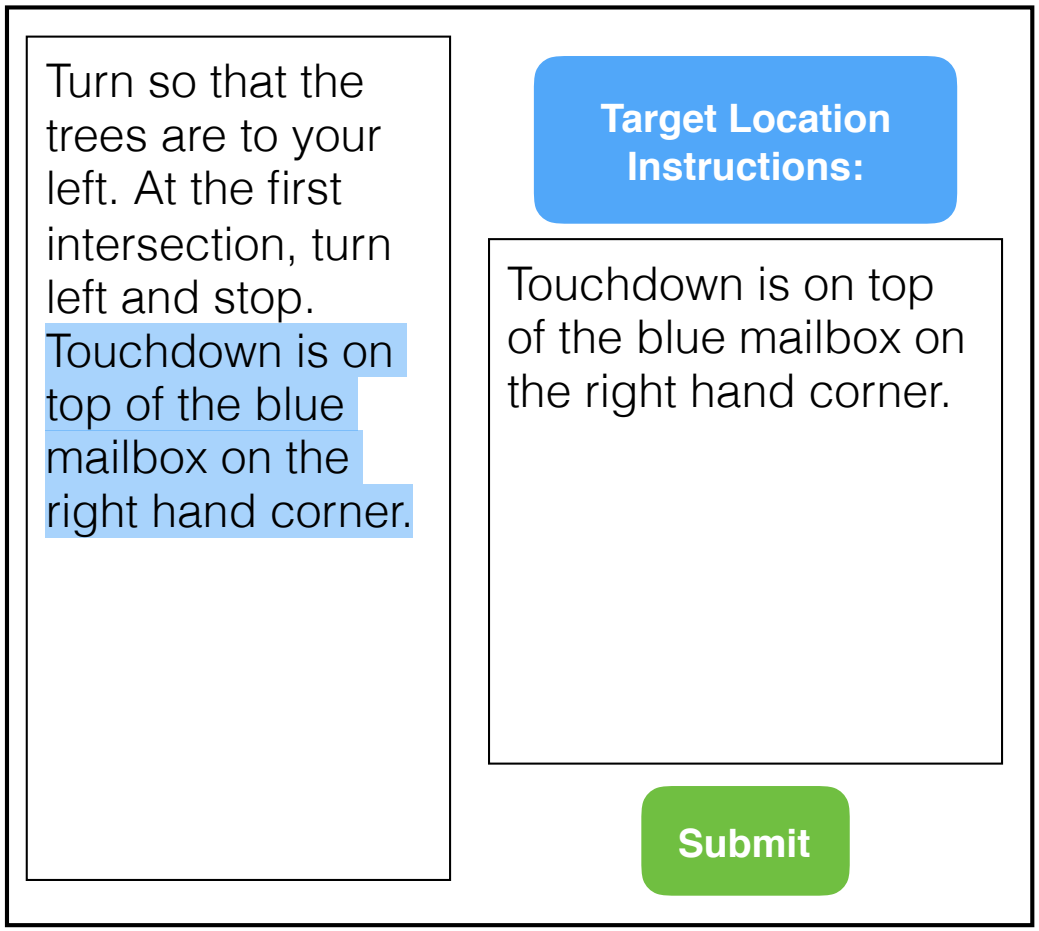
Turn so that the trees are to your left. At the first intersection, turn left and stop. Touchdown is on top of the blue mailbox on the right hand corner.

You Found Touchdown!

Remaining Attempts: 2

Task IV: Task Segmentation

- Segment the text to the two tasks: navigation and SDR
- Segments may overlap



What Did We Get?

- Over 200 people wrote and validated instructions
- Collected 9,326 examples, split to 6,526/1,391/1,409 for train/dev/test

Analysis

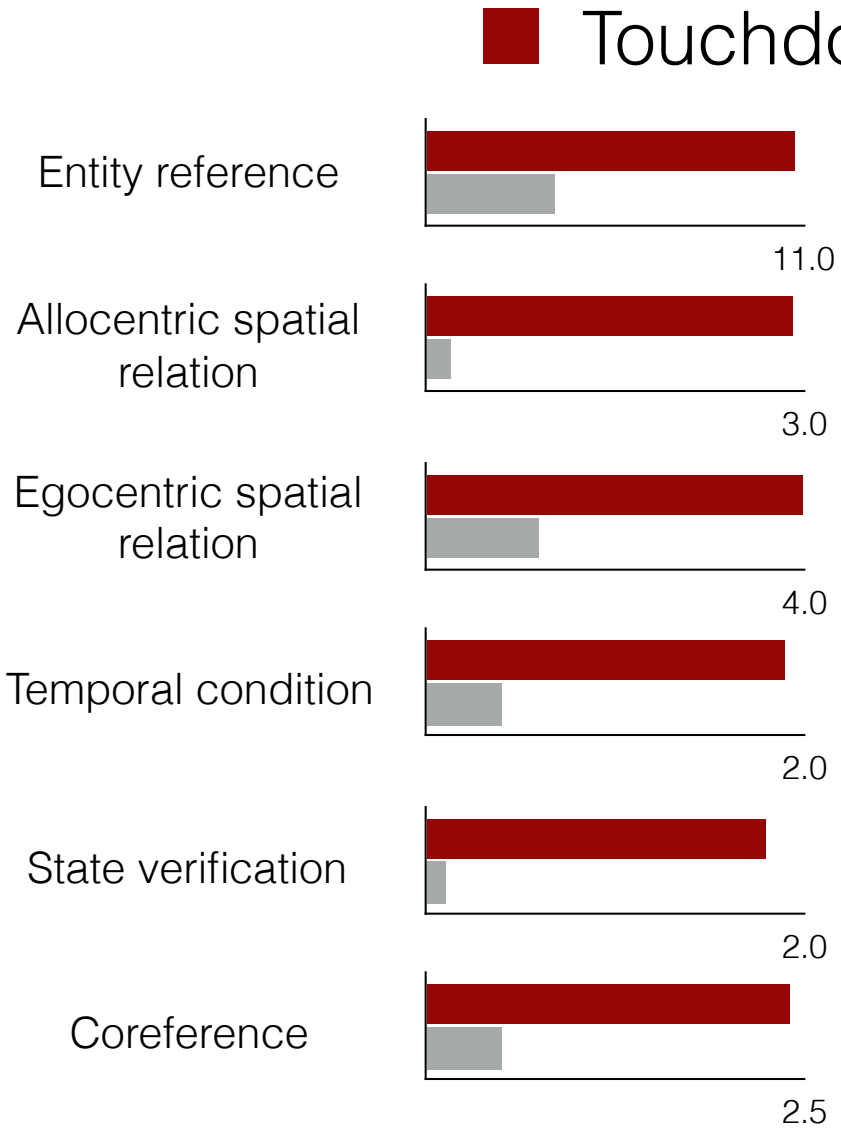
- Average length is 108 tokens on average
  - 89.6 for navigation, compared to 29.3 in R2R
  - 29.8 for SDR, compared to 8.5 in Google RefExp and 4.4 in ReferItGame
- Relatively large vocabulary size of 5,625, compared 3,156 for R2R
- Paths are on average 35.2 panoramas, compared to 6 in R2R

[R2R: Anderson et al. 2018; Google RefExp: Mao et al. 2016; ReferItGame: Kazemzadeh et al. 2014]

Linguistic Analysis

- Sampled 25 examples from Touchdown and R2R
- Analyzed for 11 semantic categories
- Report the mean number of instances per example (more analysis in the paper)

Linguistic Analysis



*…You'll pass **three trashcans** on your left …*

*… There is a fire hydrant, the bear is **on top** …*

*… up ahead there is some flag poles **on your right hand side** …*

*… Follow the road **until you see** a school on your right …*

*… **You should see** a small bridge ahead …*

*… a brownish colored brick building with a black fence around **it** …*

Spatial Description Resolution

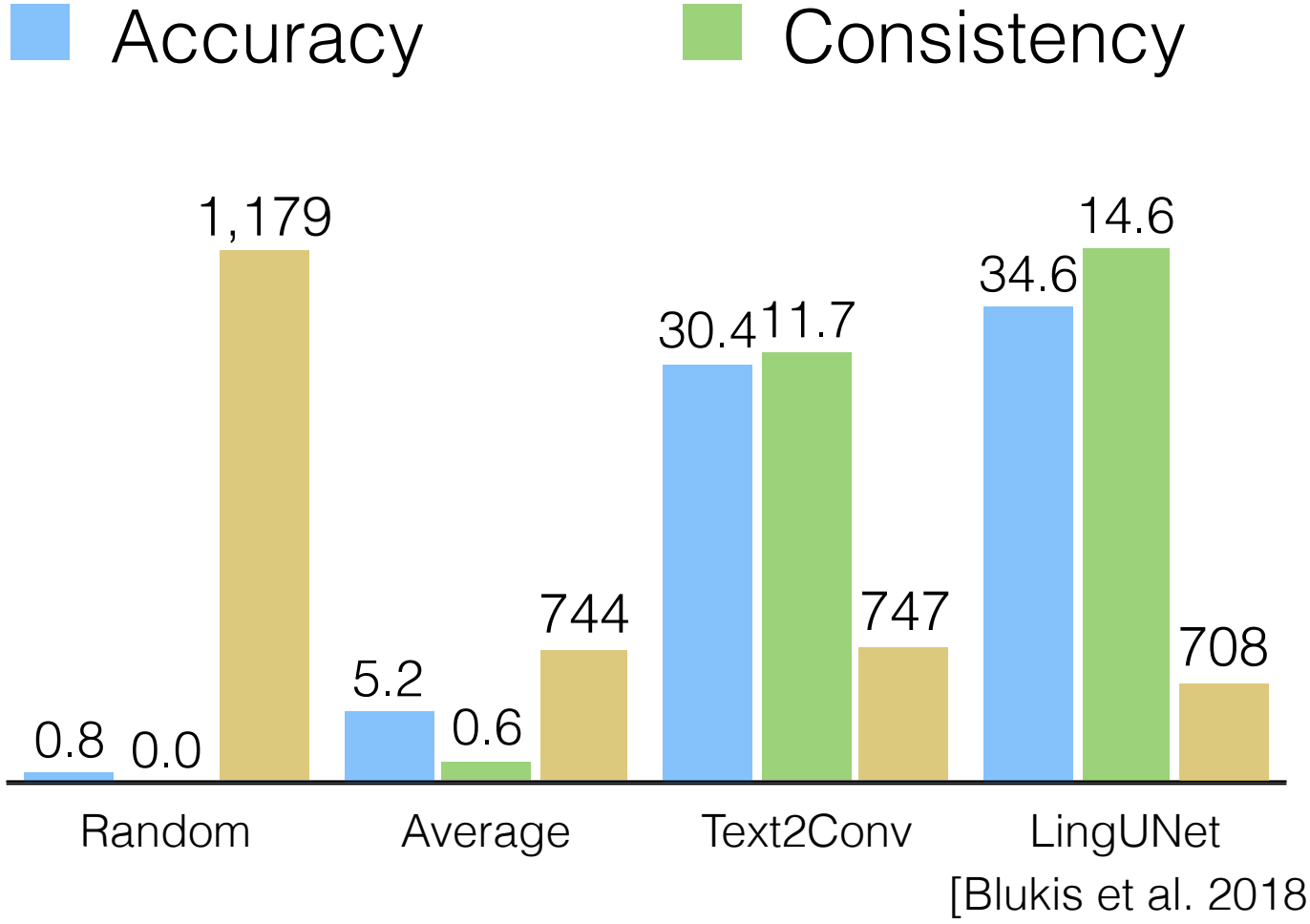
There is also a dinosaur mural to the right.
Touchdown is on the back of the dinosaur.

Where is Touchdown?

SDR Evaluation

- Accuracy: predicting the position close enough to the gold position (threshold: 80px)
- Consistency: consider a unique SDR as correct only if solved for all propagated panoramas
- Mean distance error: the distance of the predicted position from the gold position

Test Results



[Blukis et al. 2018]

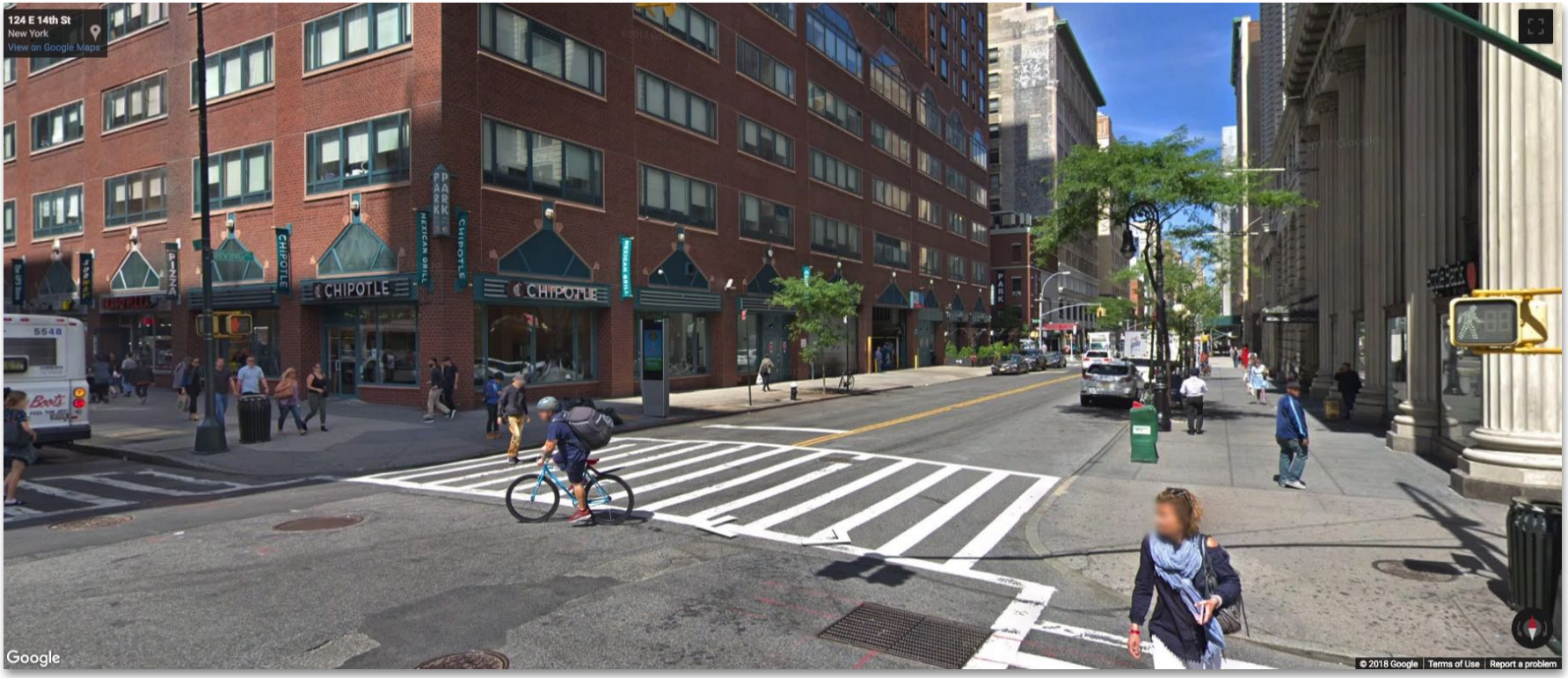
- LingUNet is able to solve some cases
- But there is a lot of room for improvement

Accuracy, consistency, and distance are not to scale

Example: LingUNet

a black doorway with red brick to the right of it, and green brick to the left of it. it has a light just above the doorway, and on that light is where you find Touchdown

Navigation



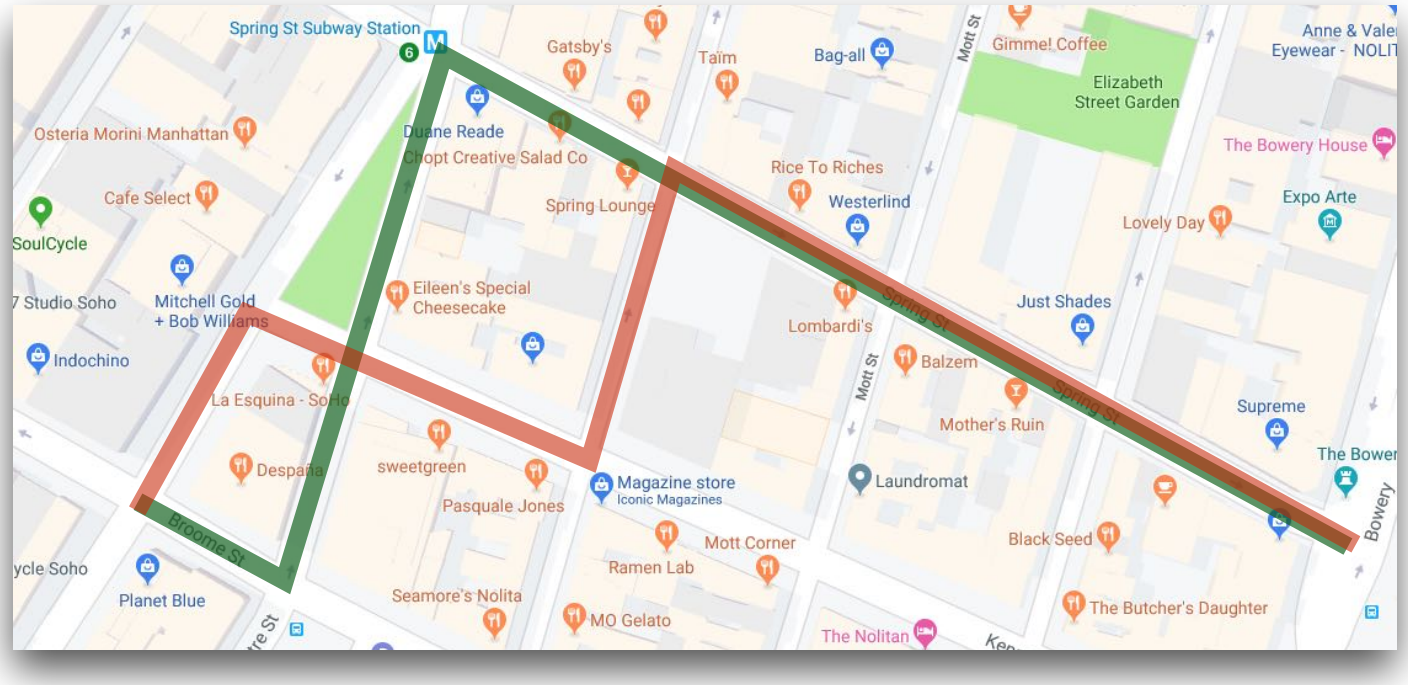
Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right.

Navigation Evaluation

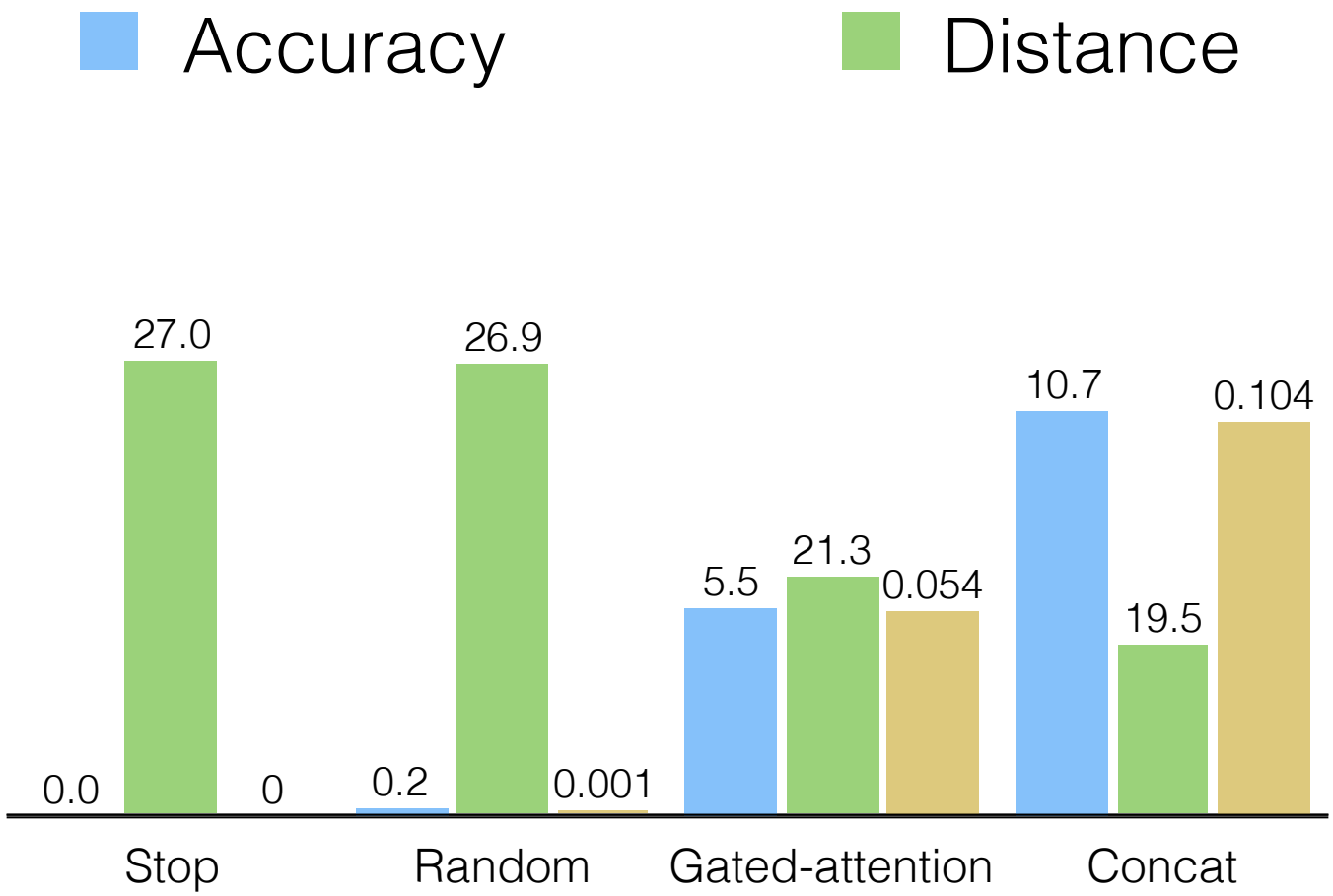
- Accuracy: stopping at the annotated goal panorama, or to one of the propagated panoramas
- Mean distance error: the shortest-path distance between the stopping position and the goal
- Success-weighted by edit distance (SED)

Success weighted by Edit Distance (SED)

- Measure edit distance between reference and prediction
- Weight success by distance
- The closer the agent is to the correct execution, success is considered better



Test Results



- Non-learning models show the task is challenging
- No model learns effectively

[Chaplot et al. 2018] [Mirowski et al. 2018]

Accuracy, distance, and SED are not to scale

Today

Understanding

NLVR

NLVR2 (*nlvr.ai*)

- Robustness to biases
- Language and reasoning diversity

Acting

Touchdown (*touchdown.ai*)

DRIF

- Real-life input
- Robotic agents

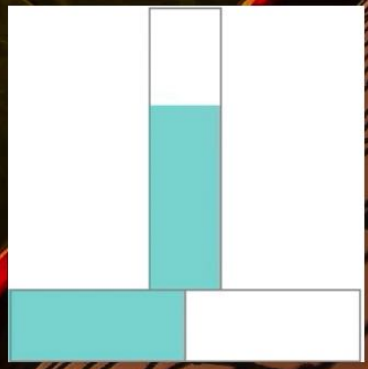
Dynamic Robot Instruction Following (DRIF)

Linear forward velocity

$$f(\text{Go towards the blue fence passing the anvil and tree on the right}, \; , \;) = \begin{bmatrix} v_t \\ STOP \\ \omega_t \end{bmatrix}$$

Angular yaw rate

after the blue bale take a right towards the small white bush before the white bush take a right and head towards the right side of the banana



Today

Understanding

NLVR

NLVR2 (*nlvr.ai*)

Alane Suhr

Stephanie Zhou
(now UMD)

Acting

Touchdown
(*touchdown.ai*)

Howard Chen

DRIF

Valts Blukis

Facebook AI Research

Google AI

Today

Understanding

NLVR

NLVR2 (*nlvr.ai*)

Alane Suhr

Stephanie Zhou
(now UMD)

- Robustness to biases
- Language and reasoning diversity

Acting

Touchdown (*touchdown.ai*)

Howard Chen

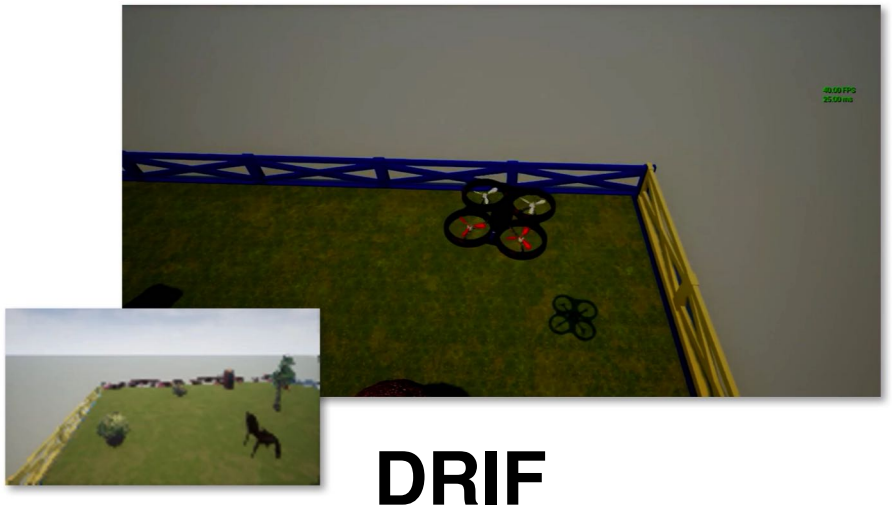
DRIF

Valts Blukis

- Real-life input
- Robotic agents

Resources: Visual Understanding to Interaction

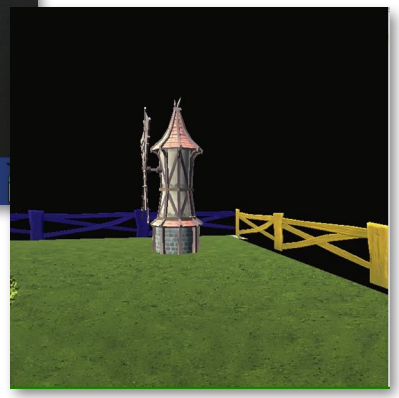
nlvr.ai



DRIF

touchdown.ai

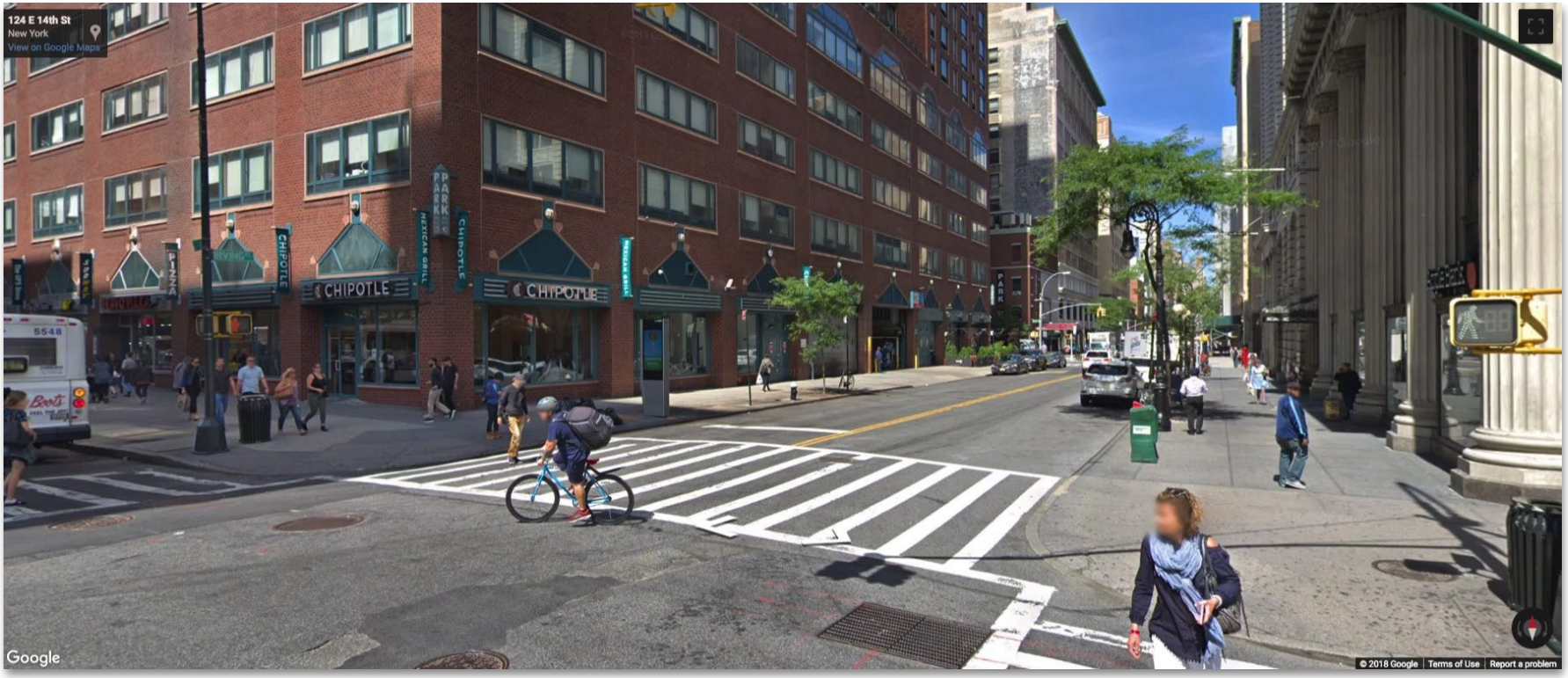
CHALET



LANI

[fin]

Example 1 Video (Backup)



Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

Back