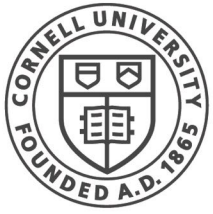
Three Challenging Research Avenues (in language and vision)

Yoav Artzi

Visual QA Workshop, CVPR 2019



Cornell CIS
Computer Science

Why Language?

- A way to outline complex and interesting visual reasoning processes
- Accessible (for data collection) and expressive
- But: diversity of reasoning relies on diversity of language

Biases and Reasoning Diversity

VQA

What is the dog carrying?

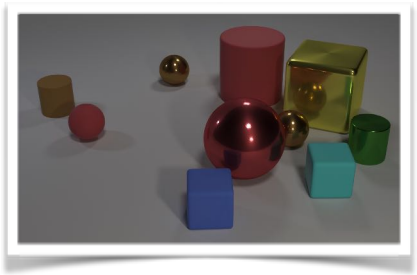
Stick

- Implicit biases
- Relatively simple language

How to stress-test our methods?

Biases and Reasoning Diversity

CLEVR



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

Yes

GQA

What color are the skis?

Black

NLVR

there are exactly three squares not touching any edge

TRUE

[CLEVR: Johnson et al. 2017; GQA: Hudson and Manning 2019]

Natural Language Visual Reasoning (NLVR)

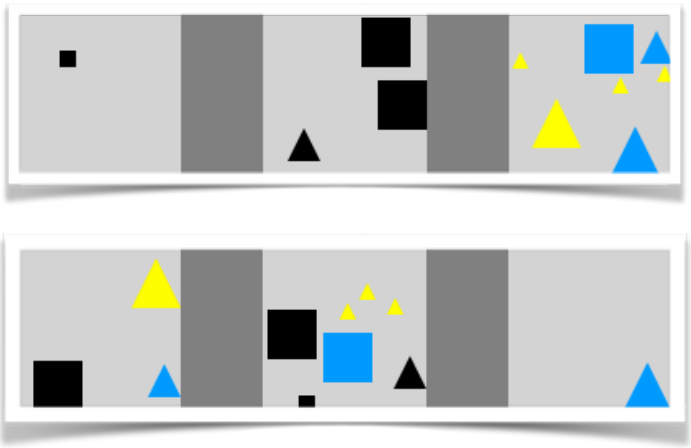
there are exactly three squares not touching any edge

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



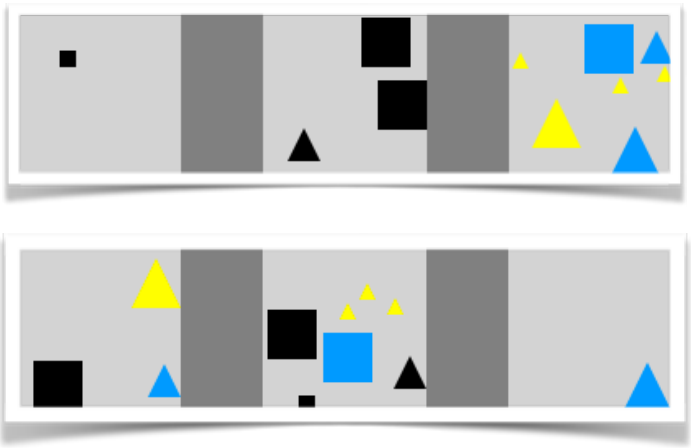
TRUE

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

FALSE

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

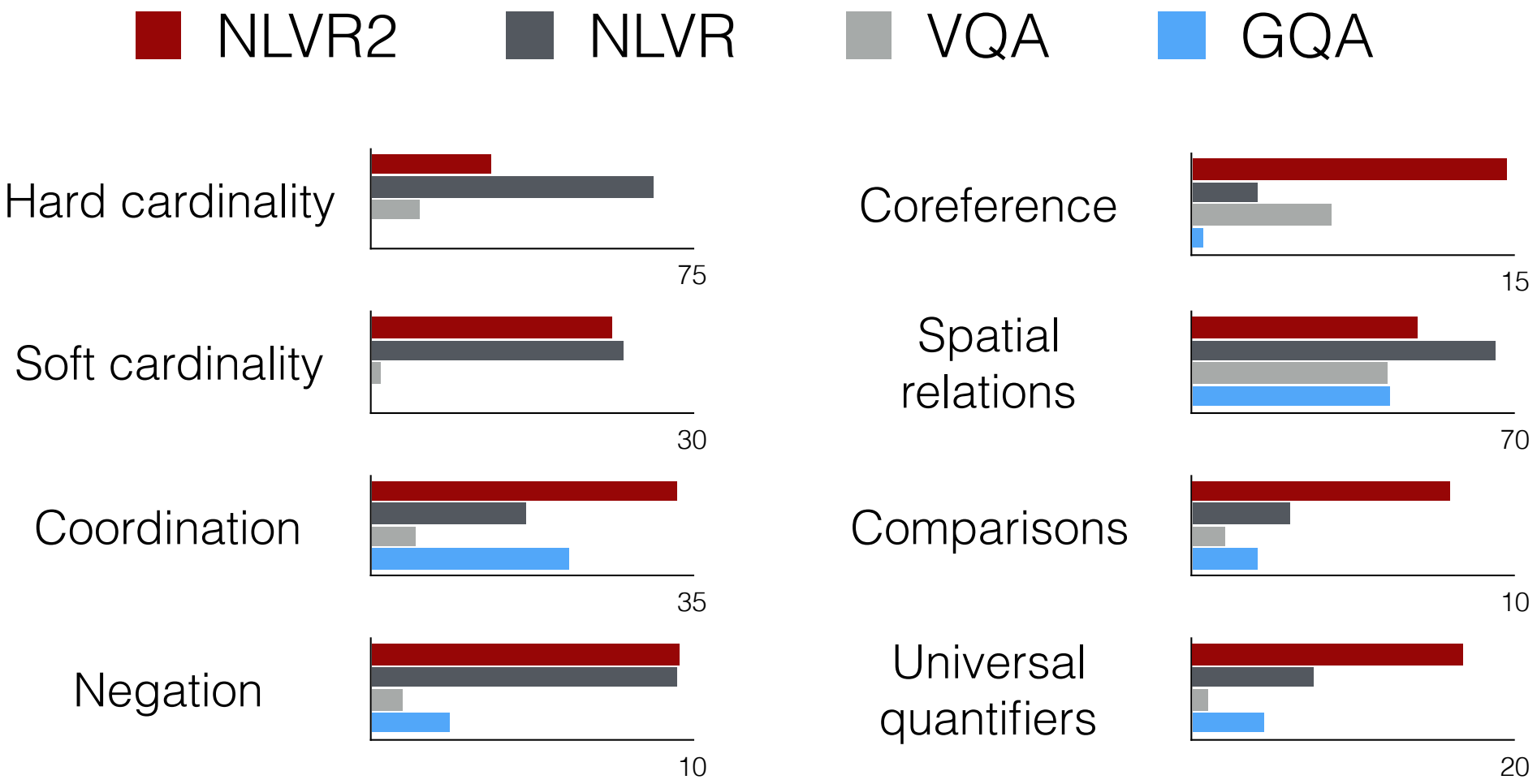
Data Collection

- Collected a new set of web images
  - Goal: *interesting* images (e.g., showing sets)
  - Aligned to 124 ImageNet synsets

Statistics

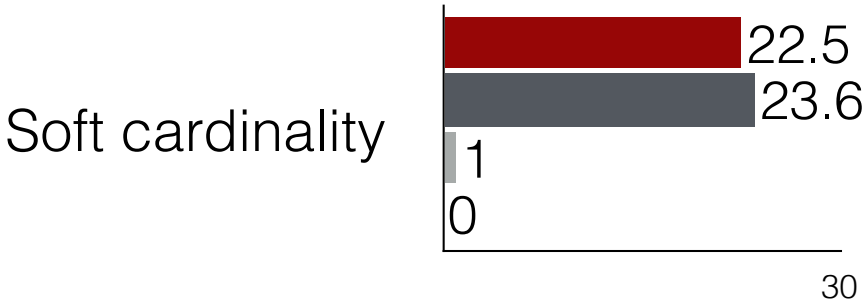
- **107,296 total examples**
  - 29,680 unique sentences
  - 127,506 unique images
  - 80% train, 20% evenly split to dev and two test sets
- **Agreement:** near perfect ($\alpha = 0.912$, $\kappa = 0.889$)
- **Average sentence length:** 14.8 tokens
- **Vocabulary size:** ~7,500 word types

Reasoning Analysis



Soft Cardinality

NLVR2 NLVR VQA GQA



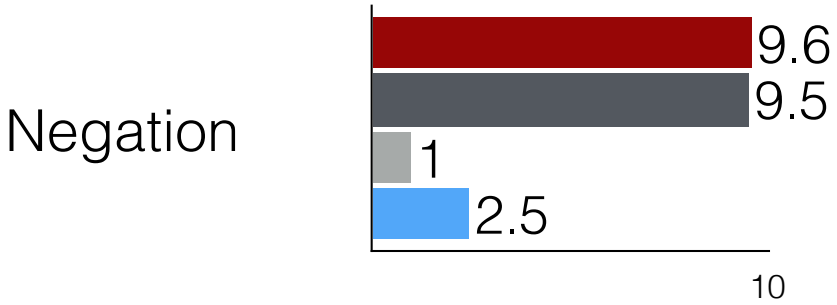
TRUE

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

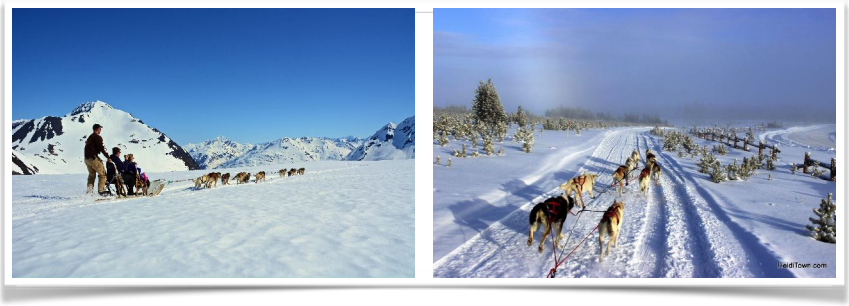
Negation

NLVR2 NLVR VQA GQA



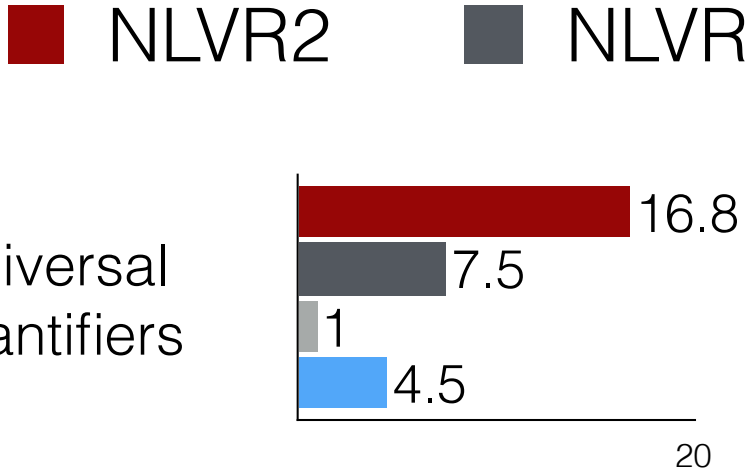
One dog sled team is moving and one is not

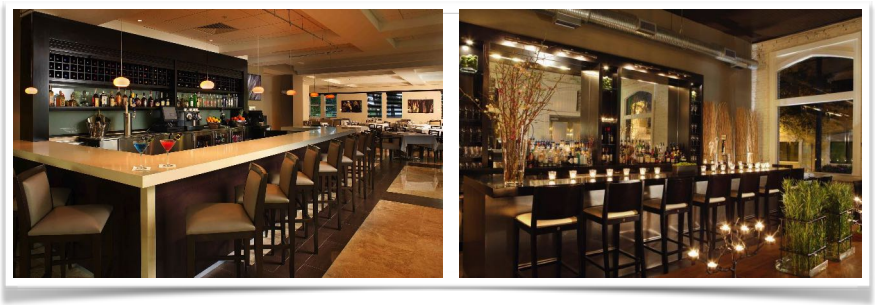
TRUE



FALSE

Universal Quantifiers





TRUE

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

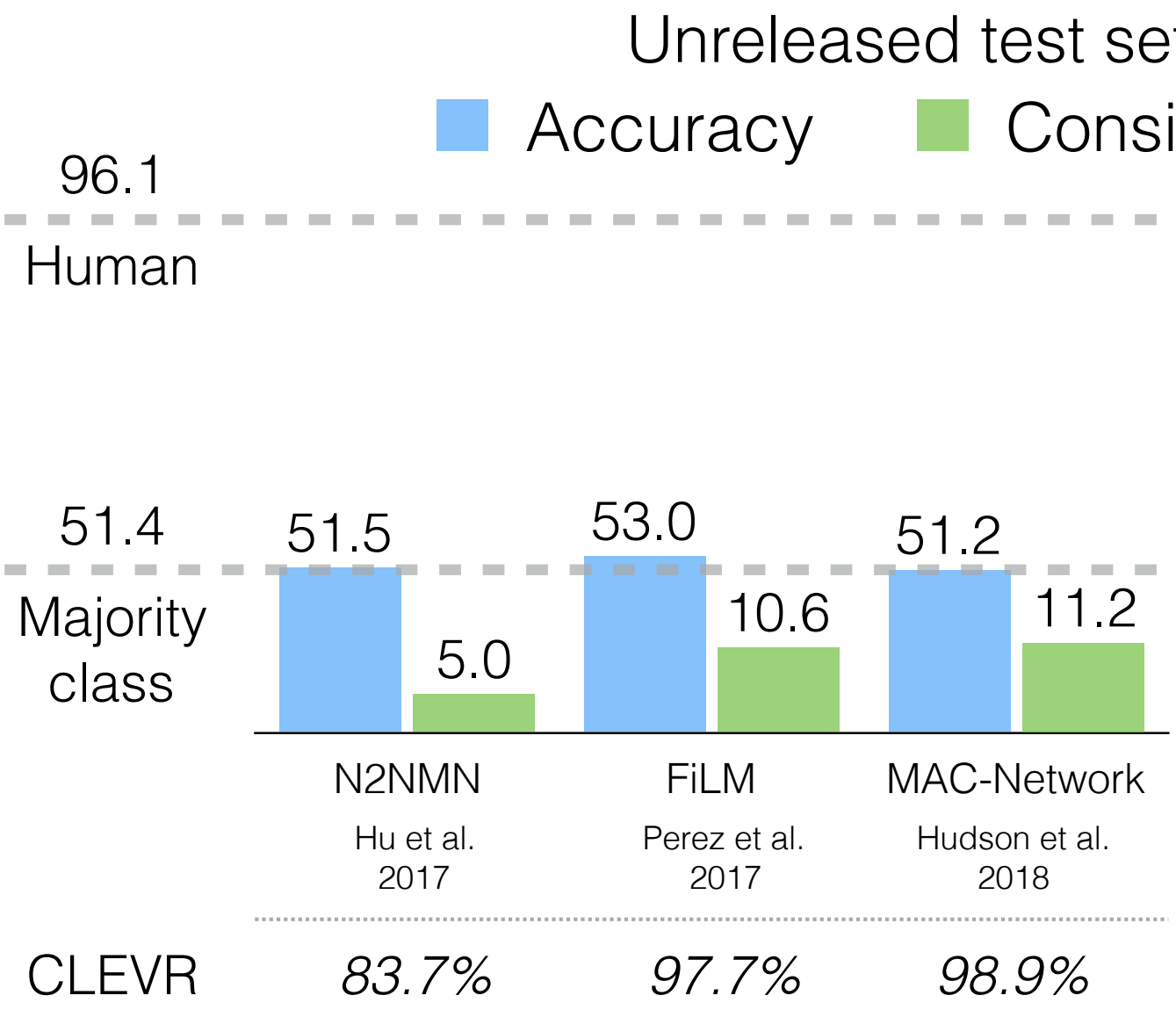
FALSE

Evaluation

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

[Goldman et al. 2018]

SOTA Visual Reasoning



- SOTA methods perform poorly
- CLEVR-NLVR2 performance mismatch

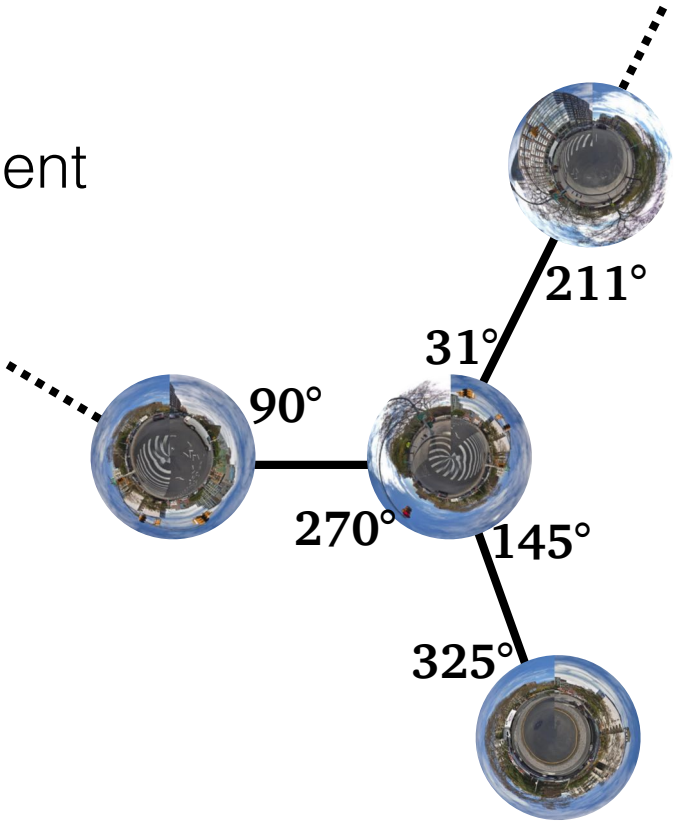
Accuracy and consistency are not to scale

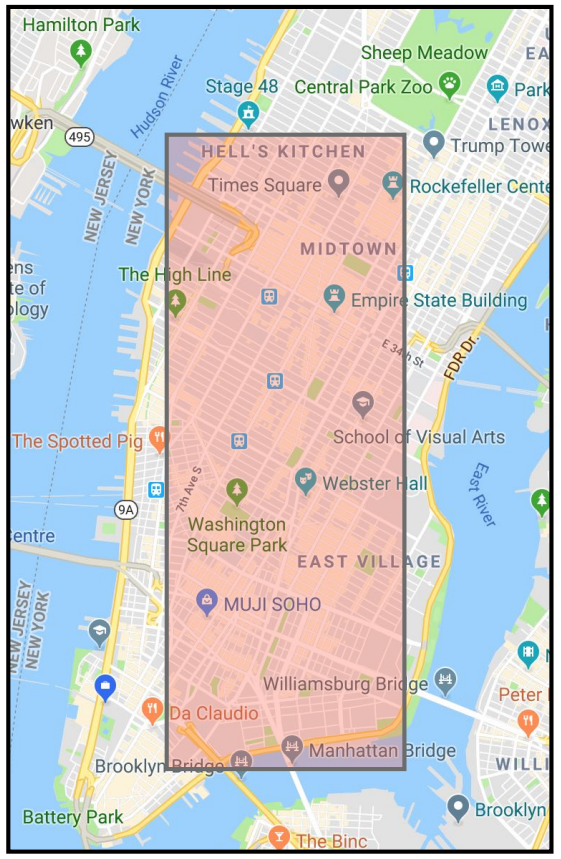

WINNER BEST MUSICAL 2009 TONY AWARD
BILLY ELLIOT THE MUSICAL
WHERE MAXIMUM PERFORMANCE LIVES
maxell
www.maxell.com
"THIS FEEL-GOOD MUSICAL IS A PHENOMENON!"
MAMMA MIA!
McDonald's
2009
TOSHIBA
TDK
3D DOUBLE FEATURE EVENT
ONLY IN THEATERS 3 WEEKS
Toy Story 2
Helped in
opponen
SPORTSMA
Pass It
Bank of America
THE LION KING
WHITE CHRISTMAS
FRANCIS P DUFFY

What kind of reasoning real-life observations entail?

The Touchdown Environment

- A large-scale visual navigation environment
- Google Street View panoramas
- 29,941 panoramas
- 61,319 edges

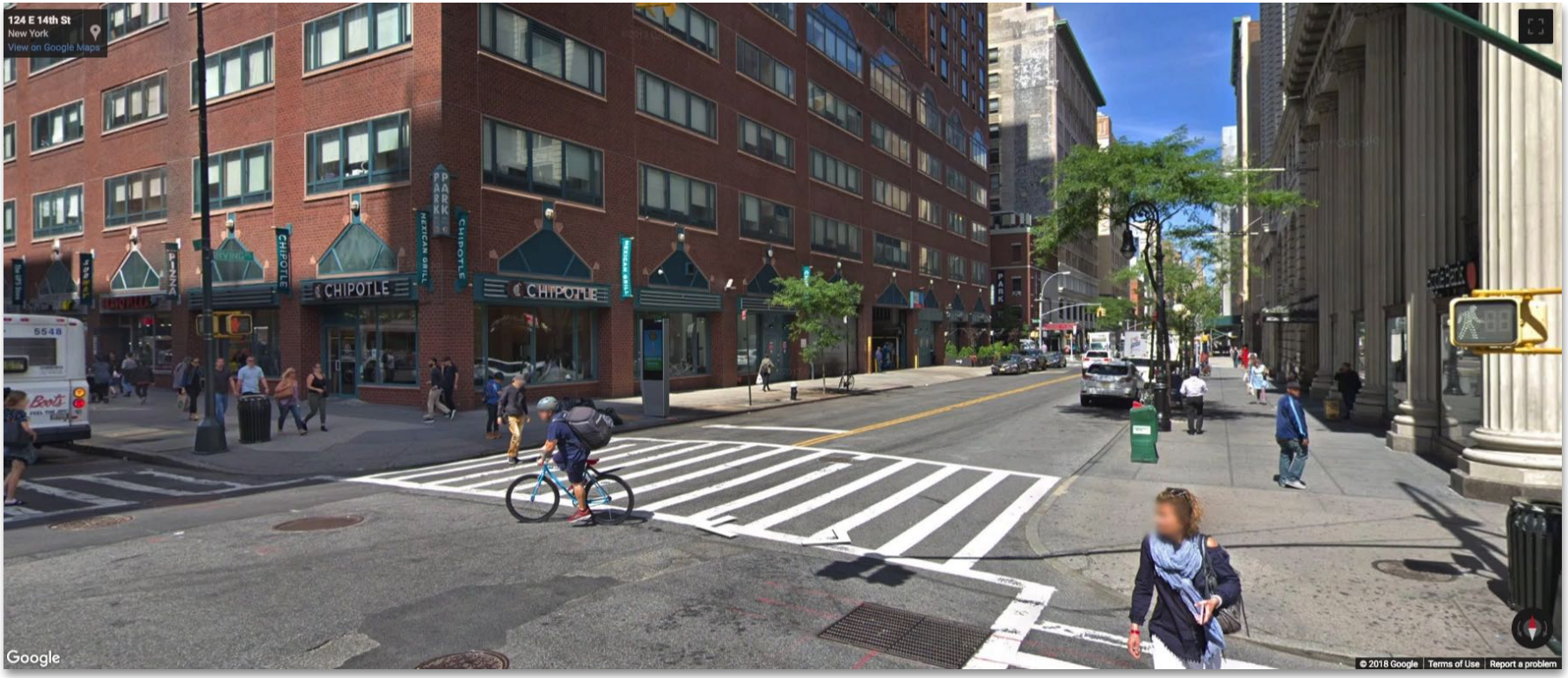




Data Collection

- Writing task: instructions to follow a path and find an object you hide — we use Touchdown
- The focused task makes the instruction more natural for the writer
- Simple validation and incentive structure
- Collected 9,326 examples

Touchdown Example



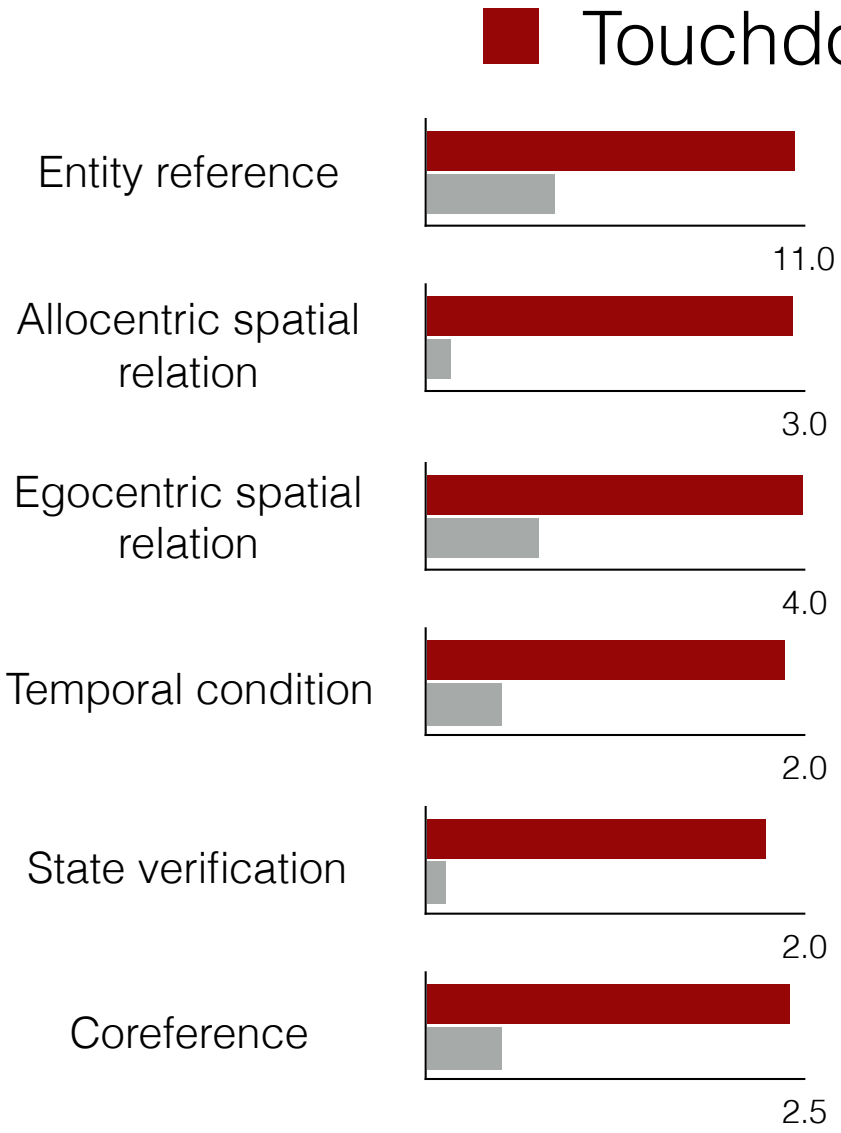
Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

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

Where is Touchdown?

Reasoning 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 …*

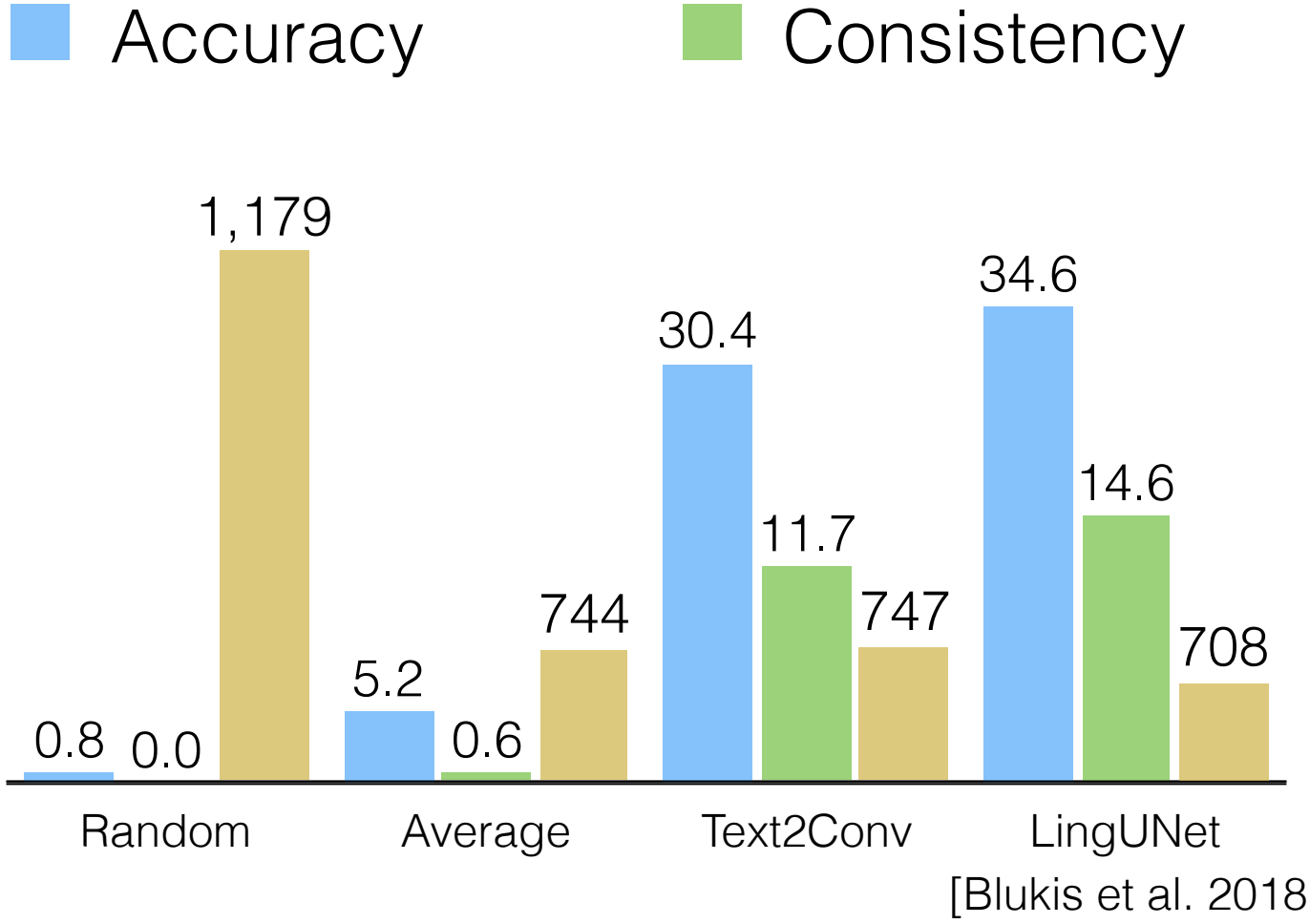
*… **You should see** a small bridge ahead …*

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

Spatial Description Resolution 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

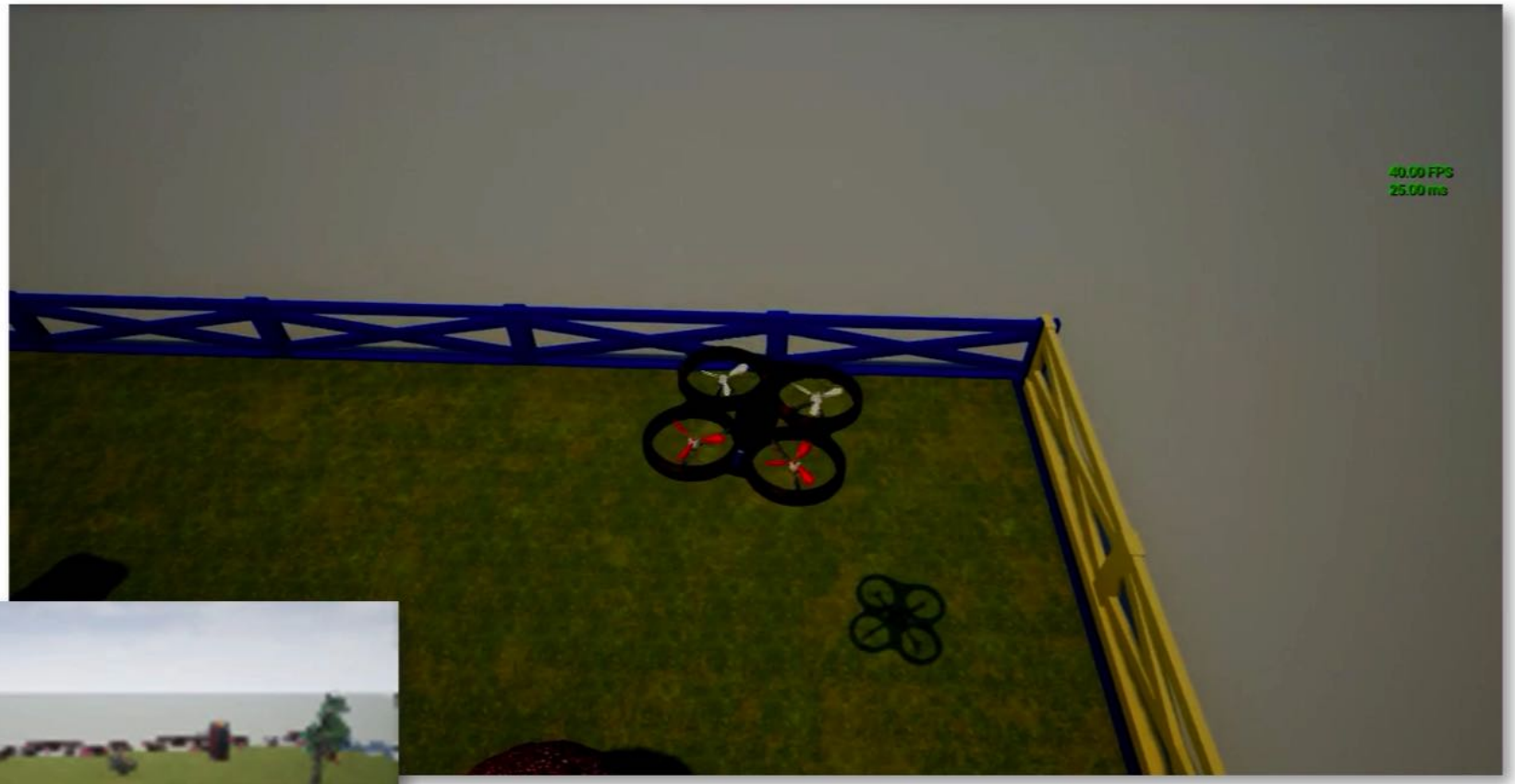
What kind of reasoning real-life observations entail?

Poster on Thursday!

touchdown.ai

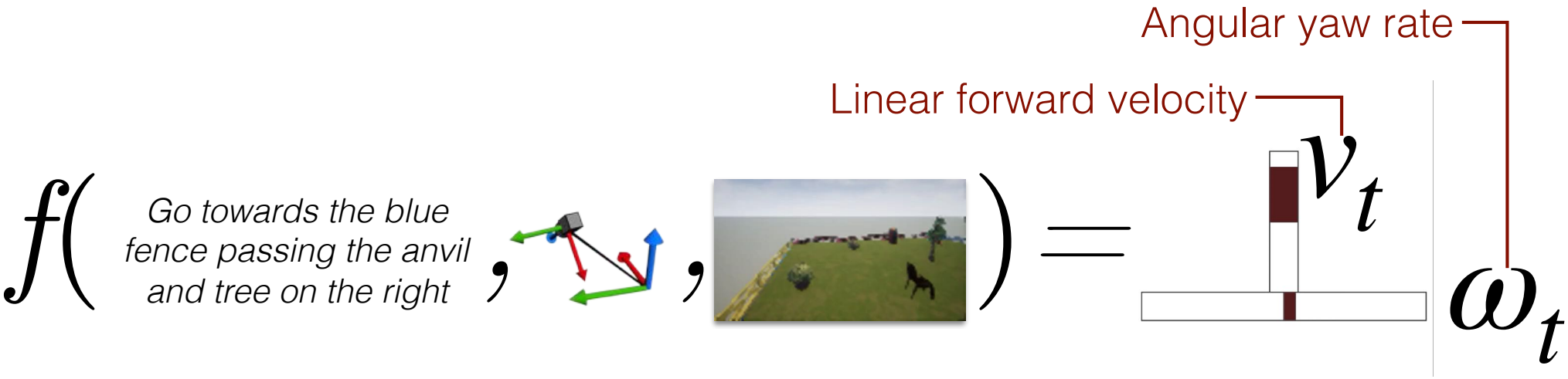
What is needed to ground to low-level actions?

Continuous Control in Visual Environments



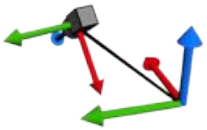
Go between the mushroom and flower chair the tree all the way up to the phone booth

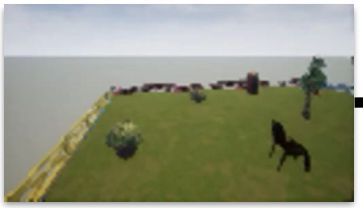
Task



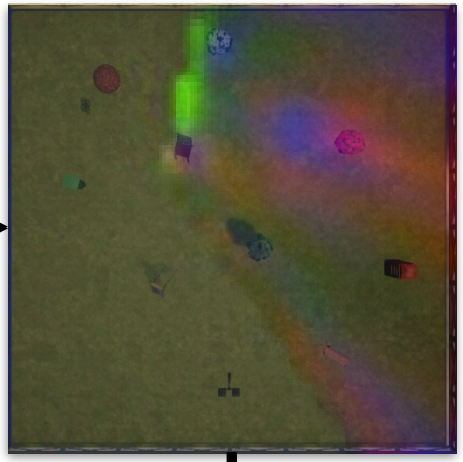
- Continuously changing observations
- Partial and accumulated observability
- Gap between low-level output and instruction

Mapping and Planning





Extraction, projection, and accumulation



Semantic Map

Go between the mushroom and flower chair the tree all the way up to the phone booth

Filtering

Grounding Map

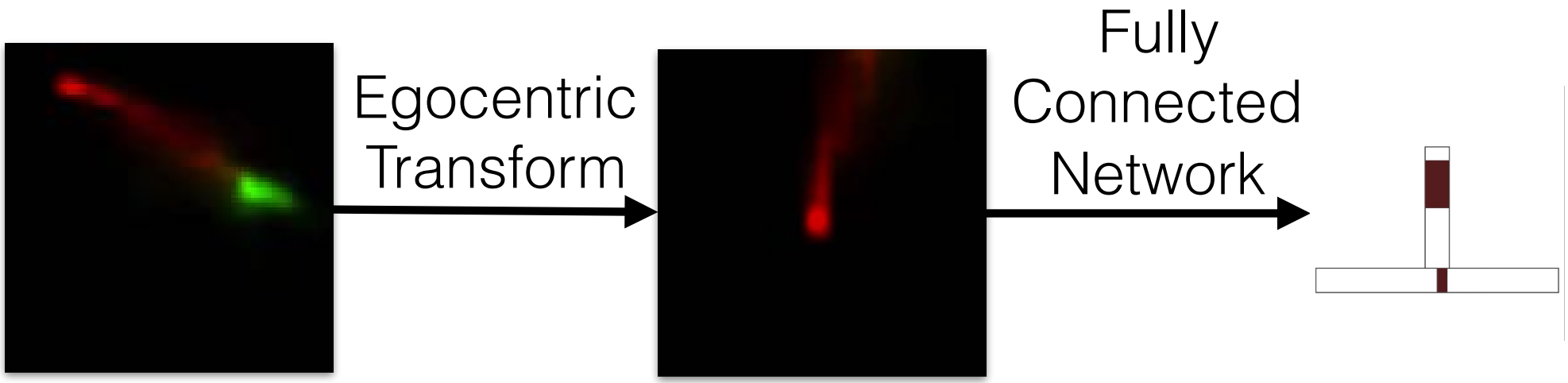
LingUNet

Trajectory Probability

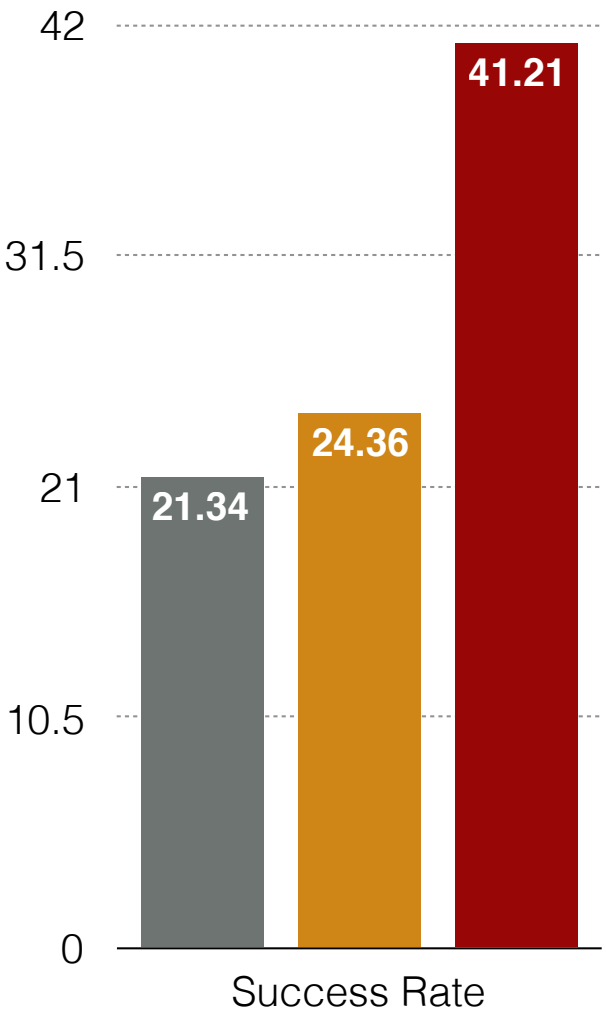
Goal Probability

Action Generation

- Relatively simple control problem without language
- Transform to agent perspective and generate configuration update



Test Results



- Chaplot et al. 2018
- Blukis et al. 2018
- Our Approach

- Explicit mapping helps performance
- Explicit planning further improves performance

Today

- Reasoning Diversity

nlvr.ai

- Real-life observations

touchdown.ai

- Realistic agents

DRIF

Alane Suhr

Stephanie Zhou

Howard Chen

Poster on Thursday!

Valts Blukis