Both Sides Now: Generating and Understanding Visually-Grounded Language

Peter Anderson

Georgia Institute of Technology

April 2019

Vision and Language

Goal: AI systems that:

Vision and Language

Goal: AI systems that:

- Communicate naturally with people

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

Image source: Lenovo

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

Image source: Lenovo

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

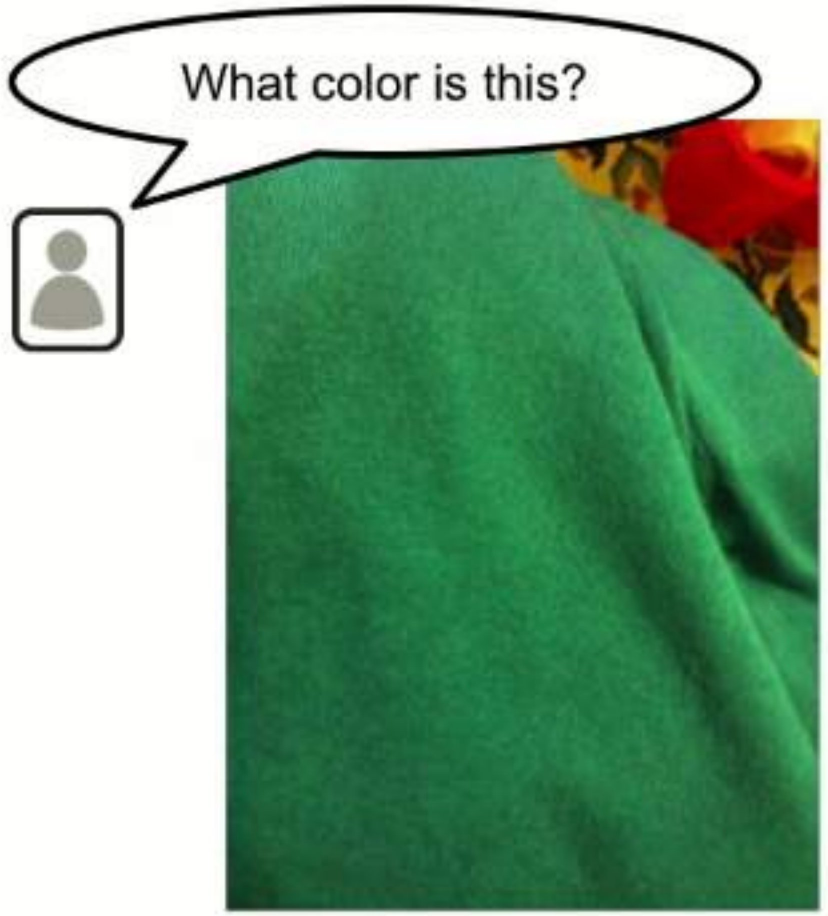
Image source: Lenovo

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Aid to the visually-impaired



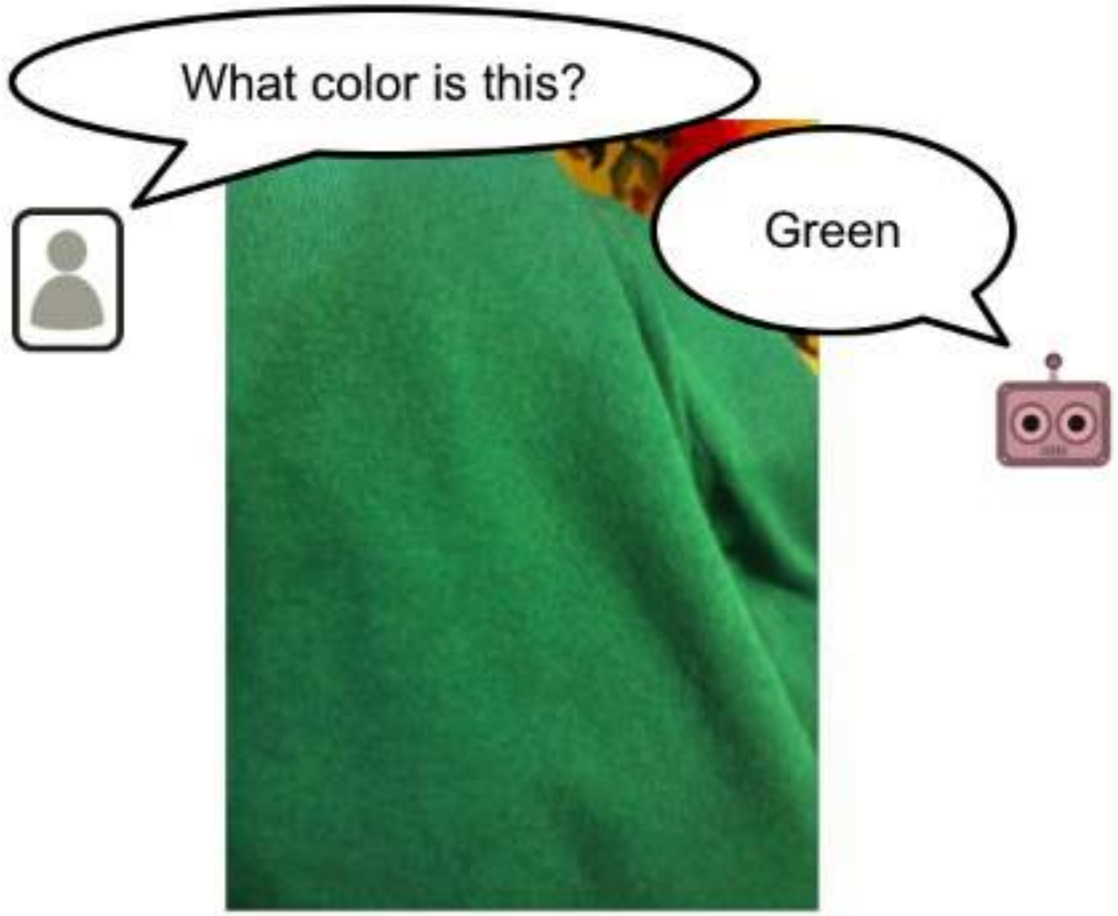
[*Gurari et al.* CVPR 2018]

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Aid to the visually-impaired



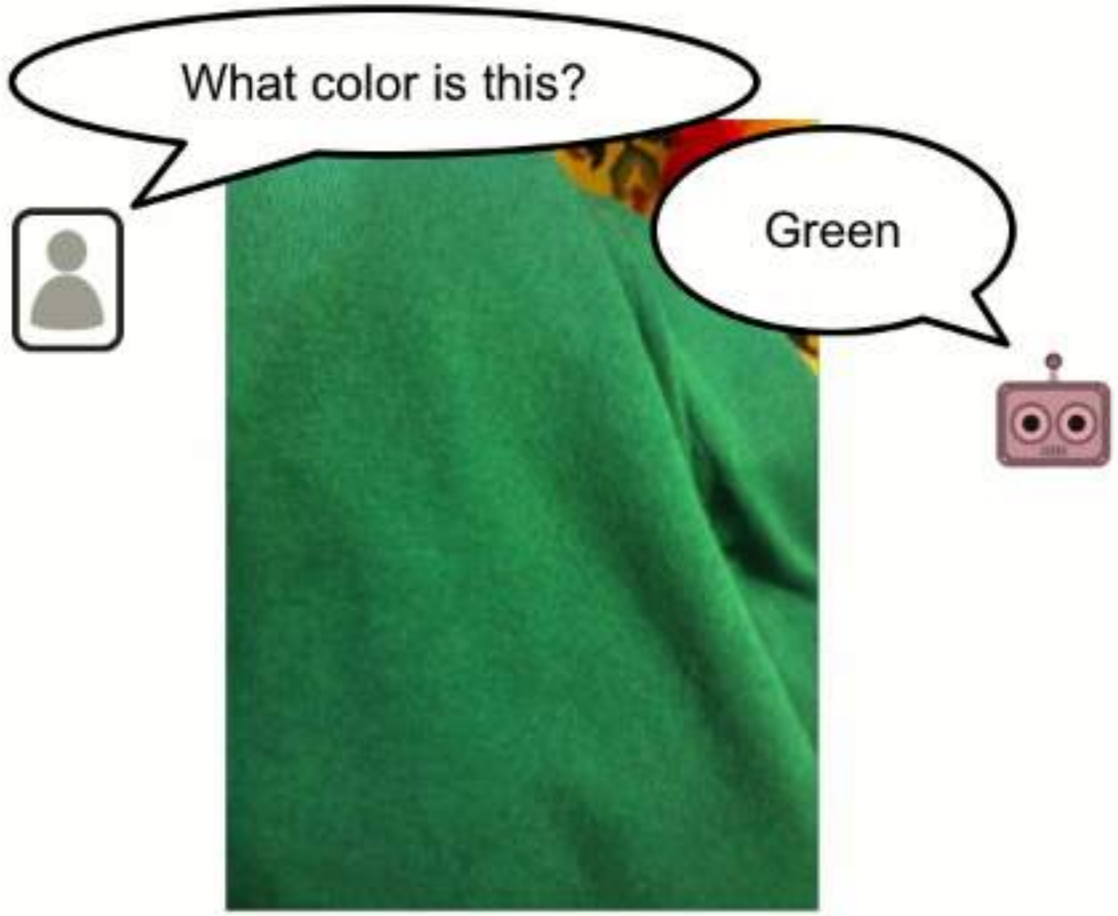
[*Gurari et al.* CVPR 2018]

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Aid to the visually-impaired



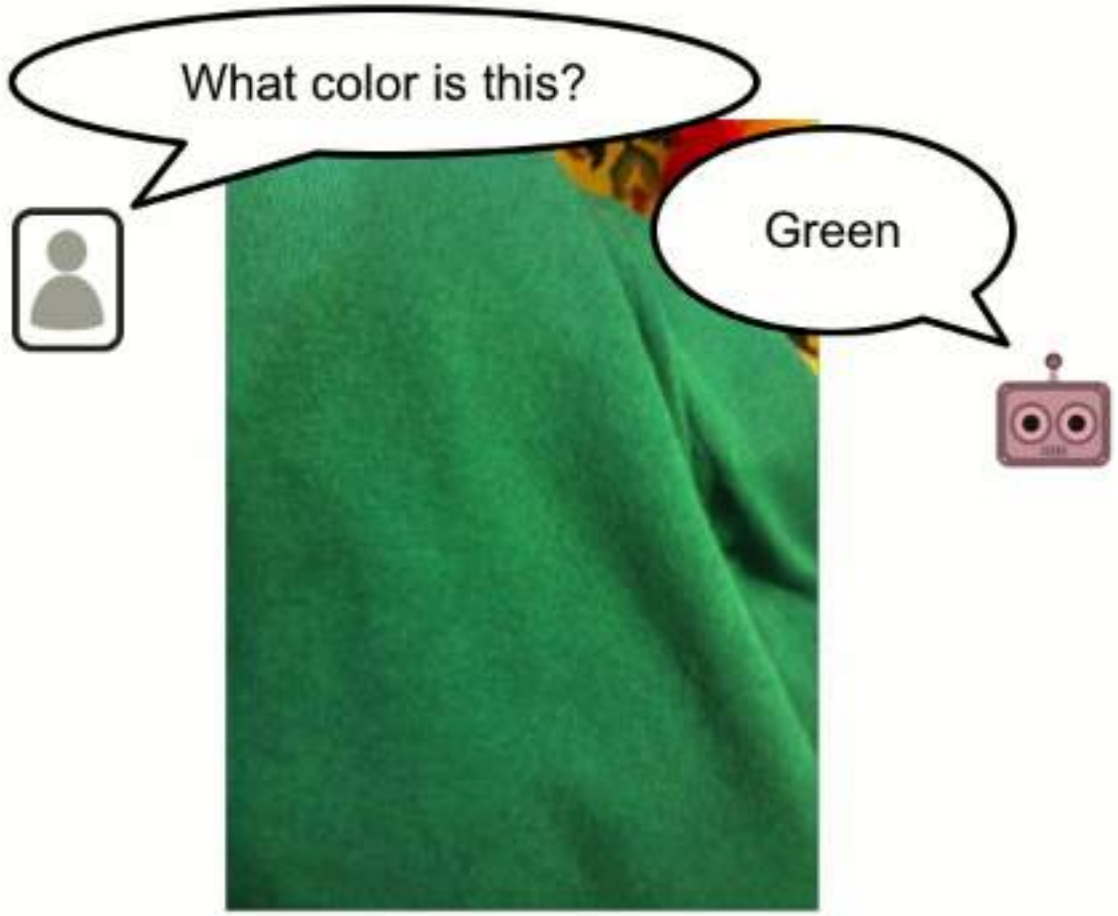
[*Gurari et al.* CVPR 2018]

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Aid to the visually-impaired



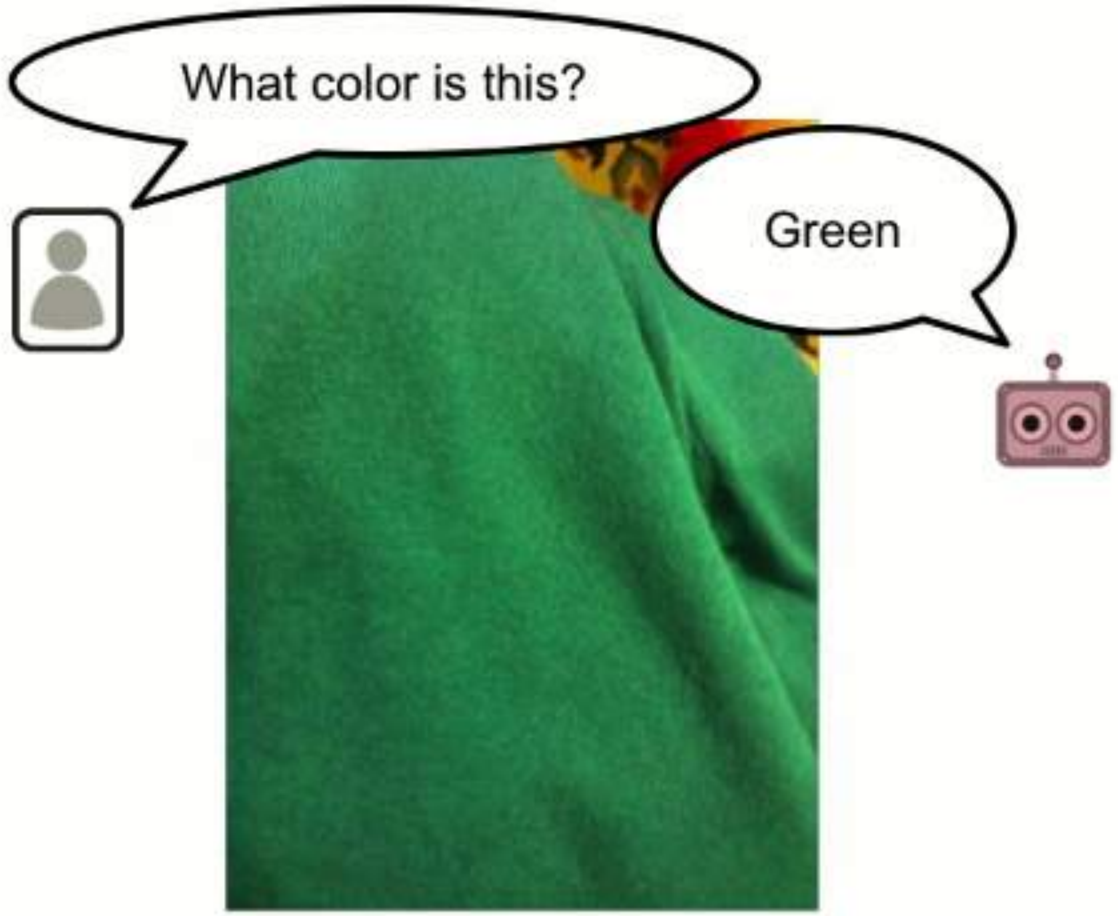
[*Gurari et al.* CVPR 2018]

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Aid to the visually-impaired



[*Gurari et al.* CVPR 2018]

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

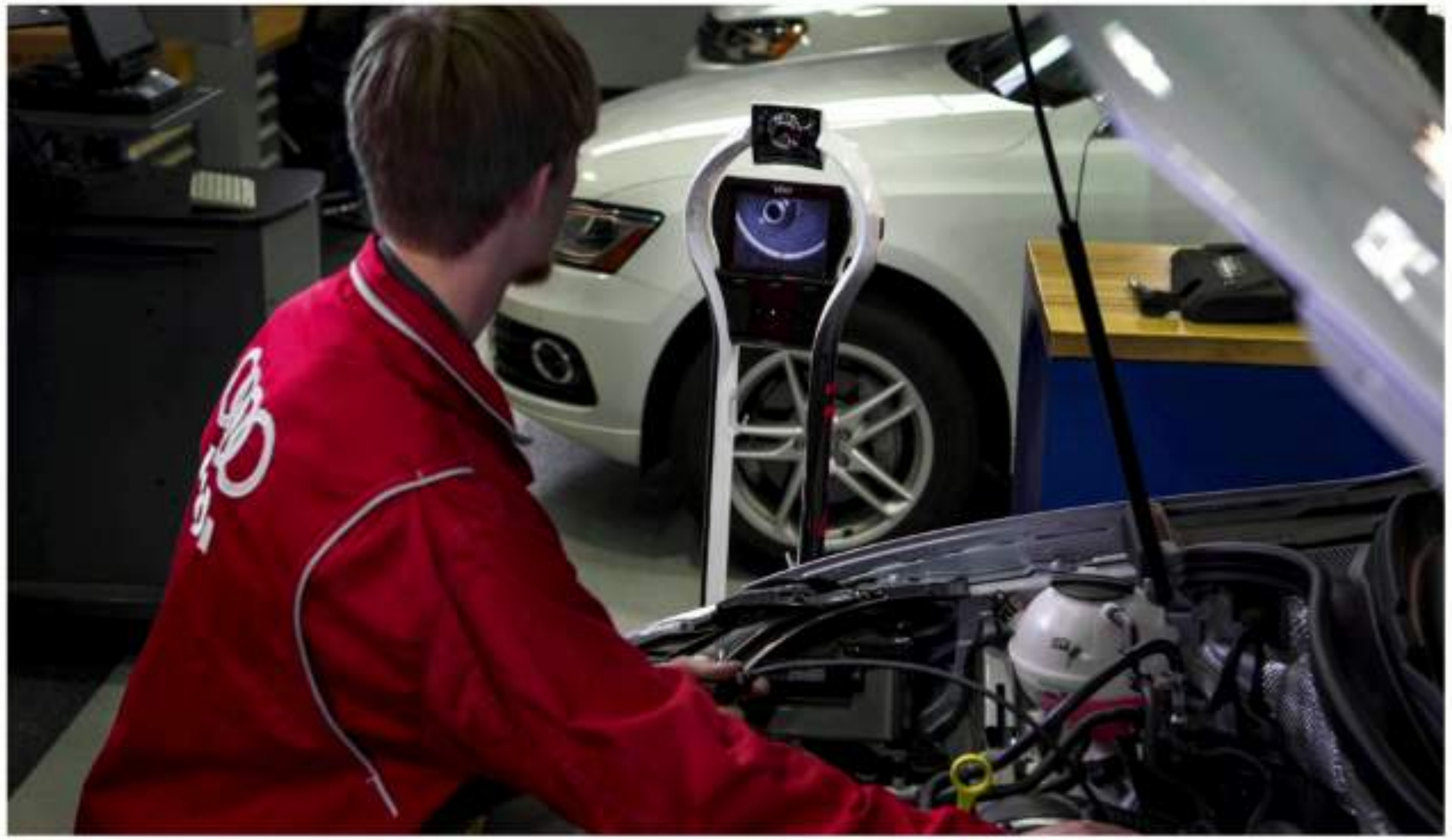


Image source: Audi

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

Image source: Audi

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

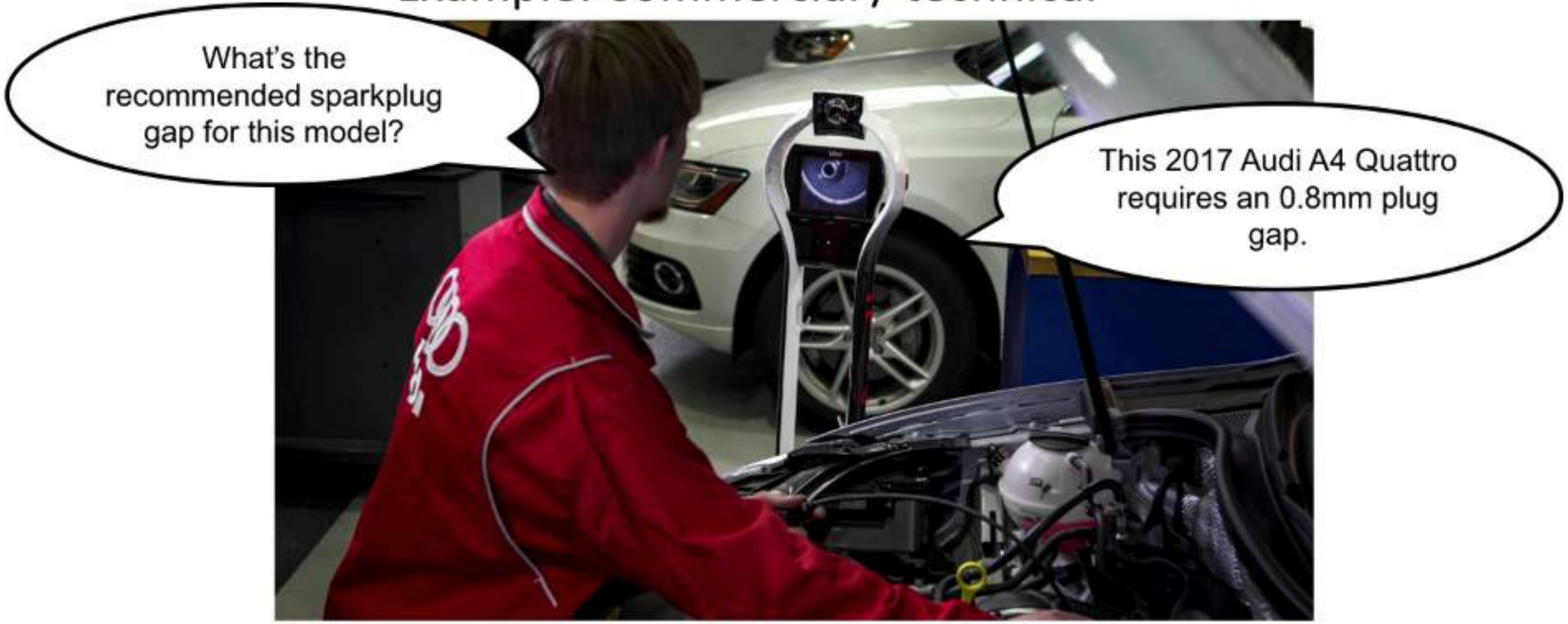


Image source: Audi

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

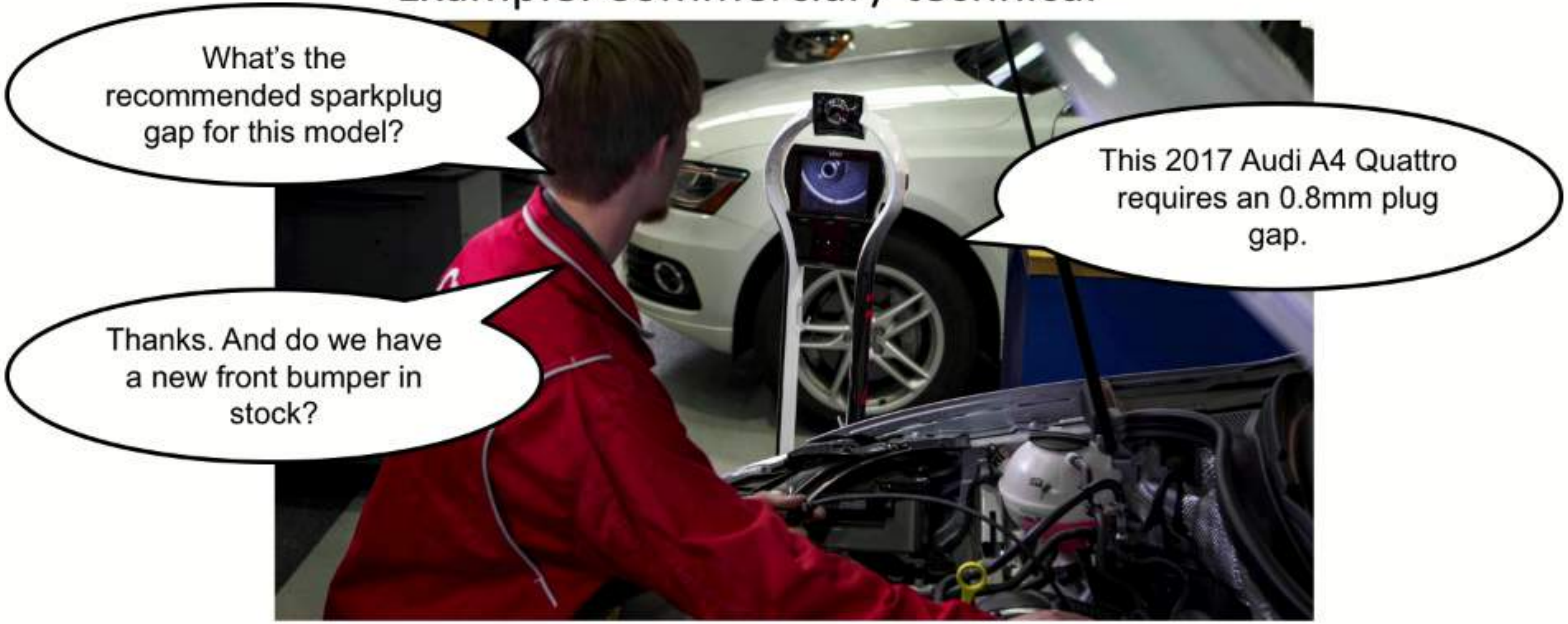


Image source: Audi

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Commercial / technical

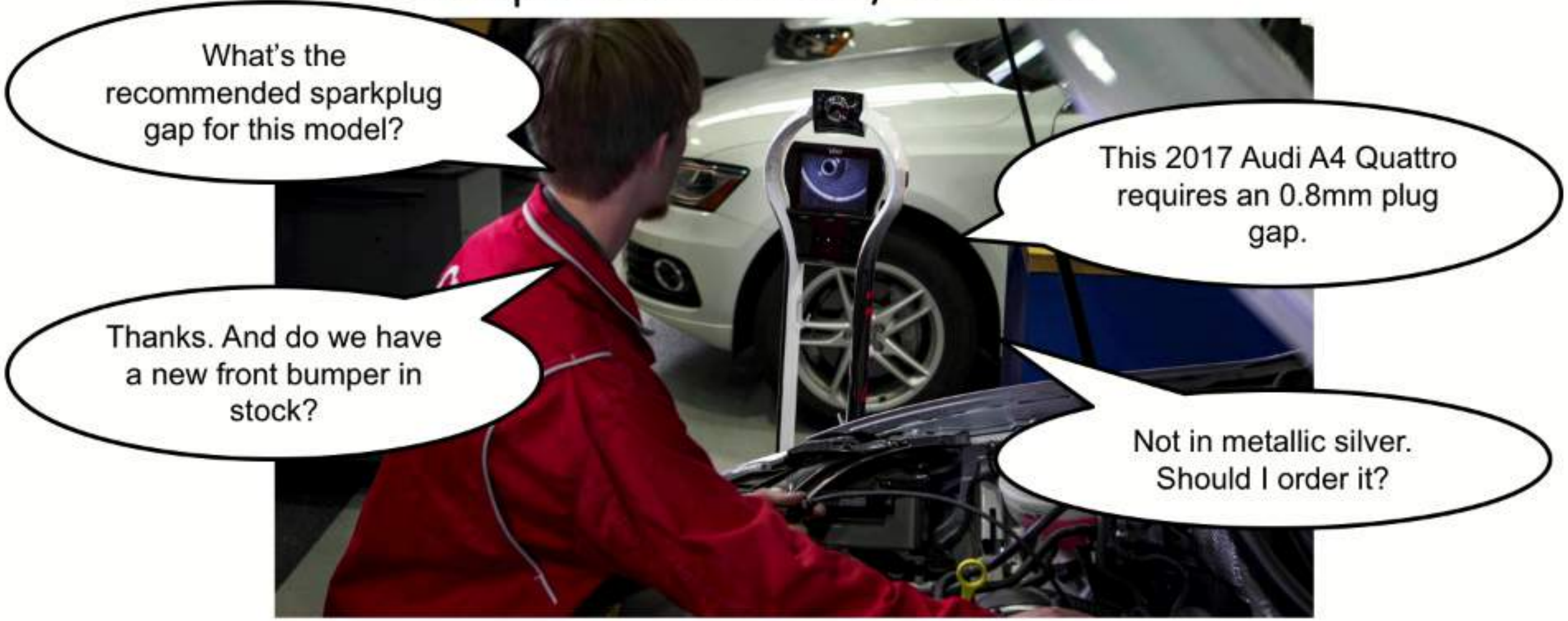


Image source: Audi

Outline

Generating Visually-Grounded Language
(Image Captioning – Novel Object Captioning)

Understanding Visually-Grounded Language
(Vision-and-Language Navigation)

Future Work

Image Captioning

- The fundamental capability to describe what is seen.

Image Captioning

- The fundamental capability to describe what is seen.

Input:

Image Captioning

- The fundamental capability to describe what is seen.

Input:

Desired Output:

A man and a woman are riding an elephant in a river.

There is a large portion of pie on a platter.

[*Chen et al.* arXiv 1504.00325 2015]

Caption Evaluation

Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):

An old green car with a flame design painted on the front of it.

A photograph of a european car.

An old school car with flames.

A picture of a car parked.

A car is painted with flames on the front.

Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):

An old green car with a flame design painted on the front of it.

A photograph of a european car.

An old school car with flames.

A picture of a car parked.

A car is painted with flames on the front.

N-grams:

1: 'A', 'teal', 'green', 'car'…

2: 'A teal', 'teal green'…

3: 'A teal green'…

N-grams:

1: 'An', 'old', 'green', 'car'…

2: 'An old', 'old green'…

3: 'An old green'…

Caption Evaluation

Candidate caption:
A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):
An old green car with a flame design painted on the front of it.
A photograph of a european car.
An old school car with flames.
A picture of a car parked.
A car is painted with flames on the front.

N-grams:
1: 'A', 'teal', 'green', 'car'…
2: 'A teal', 'teal green'…
3: 'A teal green'…

N-grams:
1: 'An', 'old', 'green', 'car'…
2: 'An old', 'old green'…
3: 'An old green'…

N-gram Similarity Score

e.g. CIDEr, BLEU, Meteor, Rouge metrics

[*Vedantam et al.* CVPR 2015, *Papineni et al.* ACL 2002, *Denkowski et al.* EACL 2014, *Lin et al.* ACL 2004]

Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):

An old green car with a flame design painted on the front of it.

A photograph of a european car.

An old school car with flames.

A picture of a car parked.

A car is painted with flames on the front.

Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):

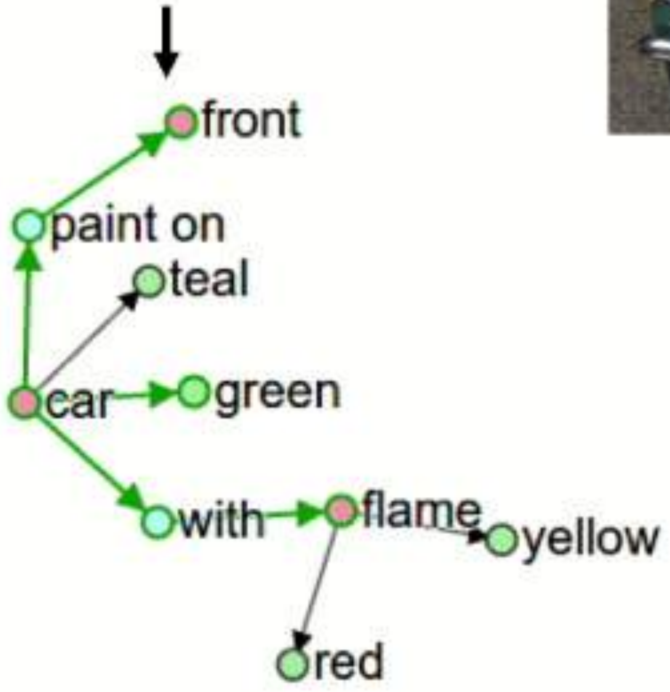
An old green car with a flame design painted on the front of it.

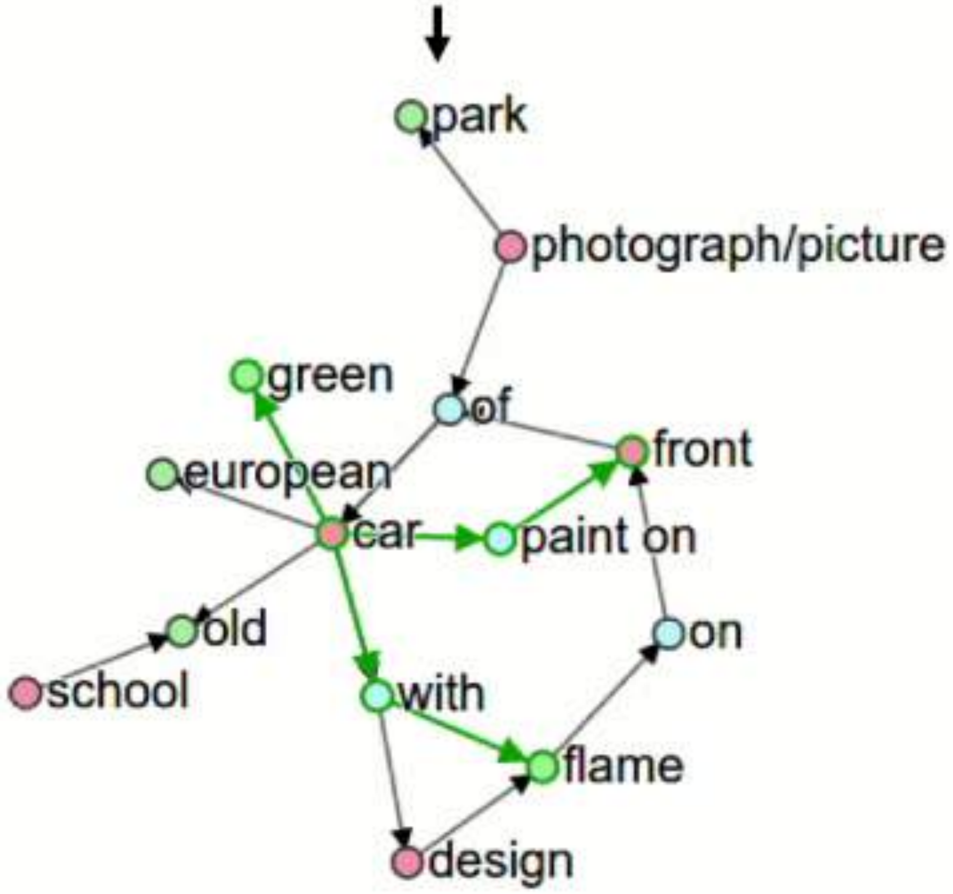
A photograph of a european car.

An old school car with flames.

A picture of a car parked.

A car is painted with flames on the front.





Caption Evaluation

Candidate caption:

A teal green car with yellow and red flames painted on the front.

Reference captions (written by people):

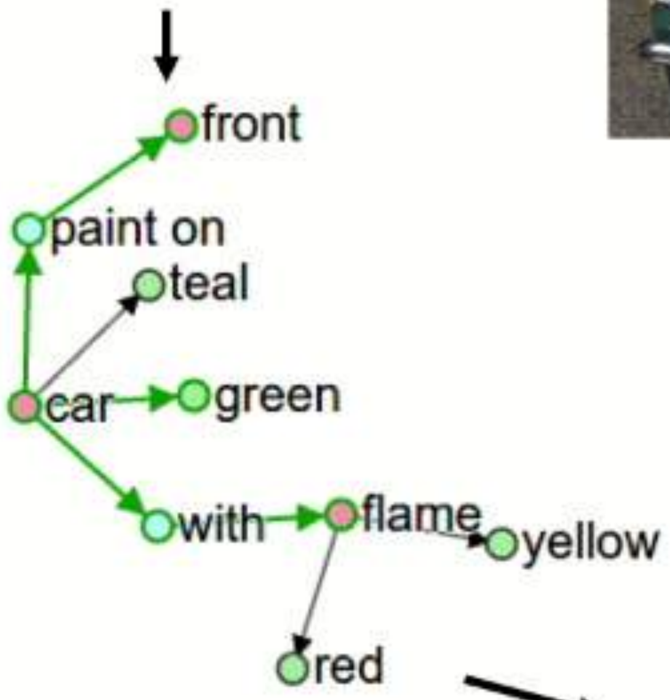
An old green car with a flame design painted on the front of it.

A photograph of a european car.

An old school car with flames.

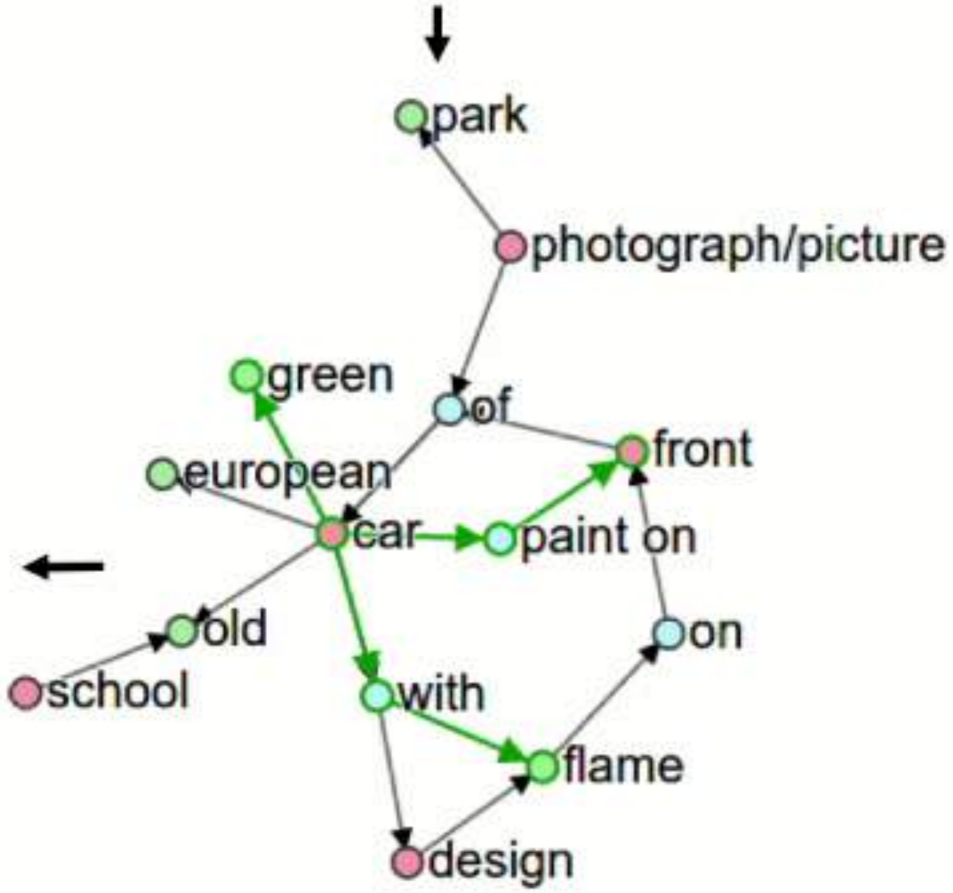
A picture of a car parked.

A car is painted with flames on the front.



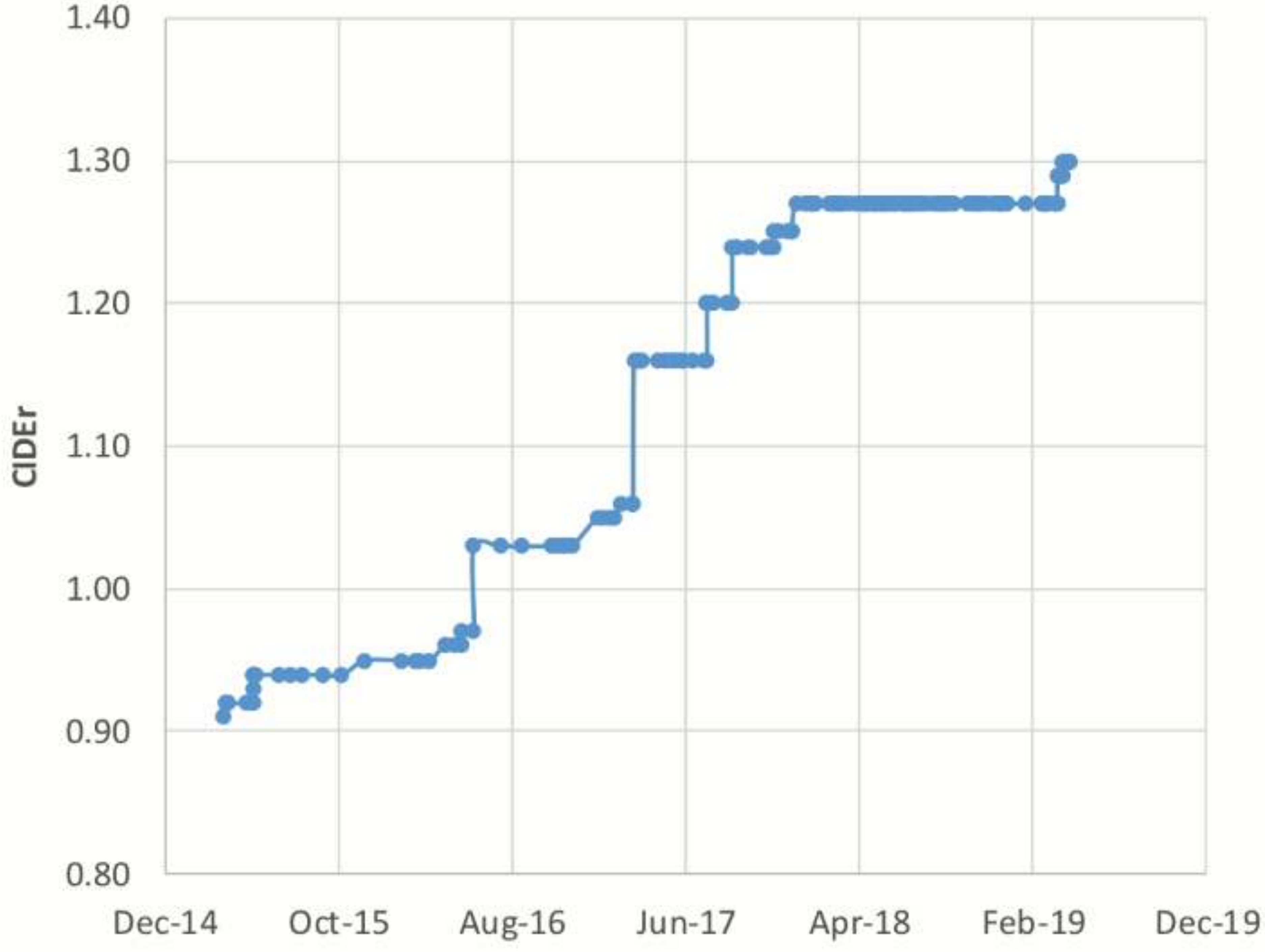
Scene-Graph Similarity Score

(SPICE metric)

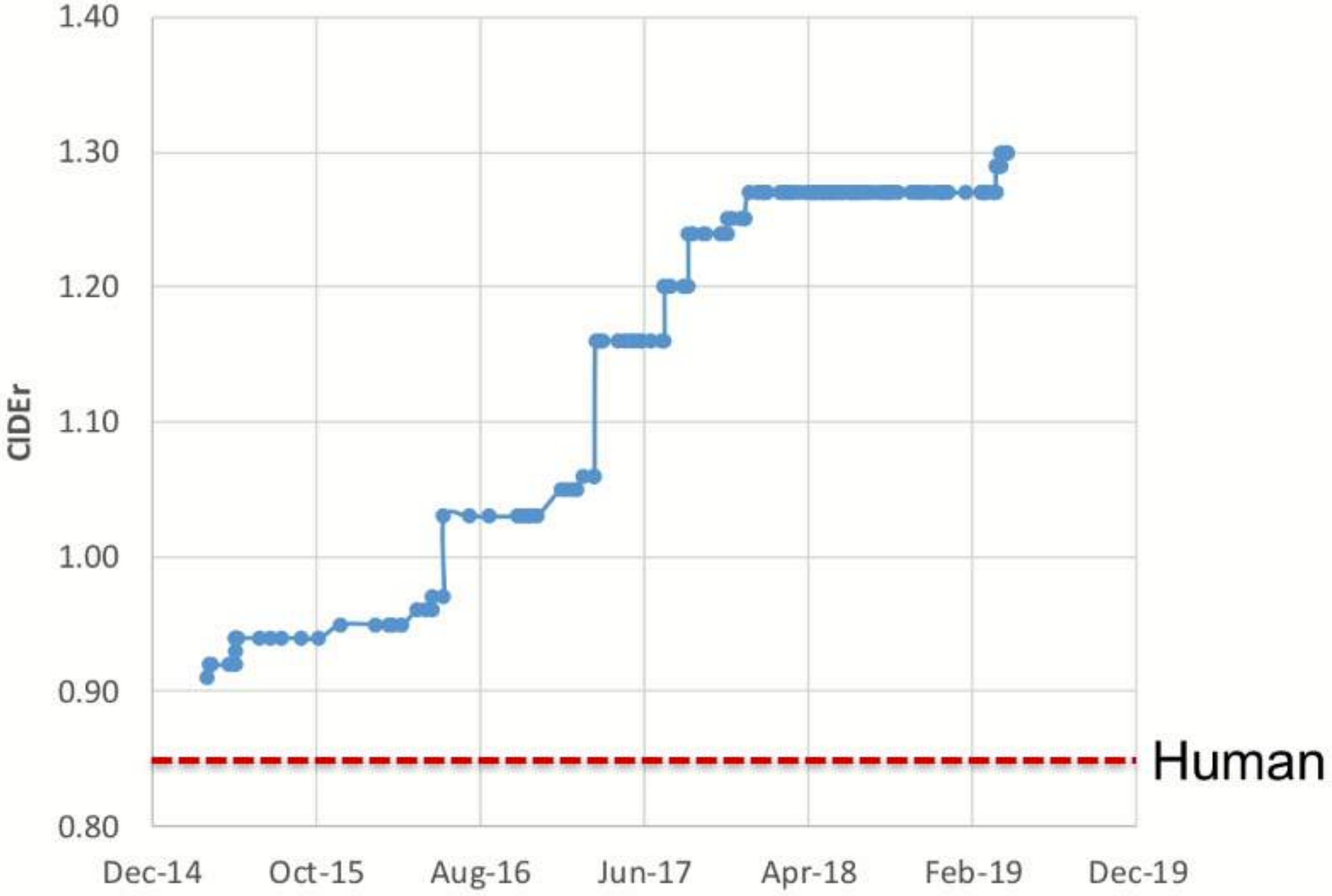


[*Anderson et al.* ECCV 2016]

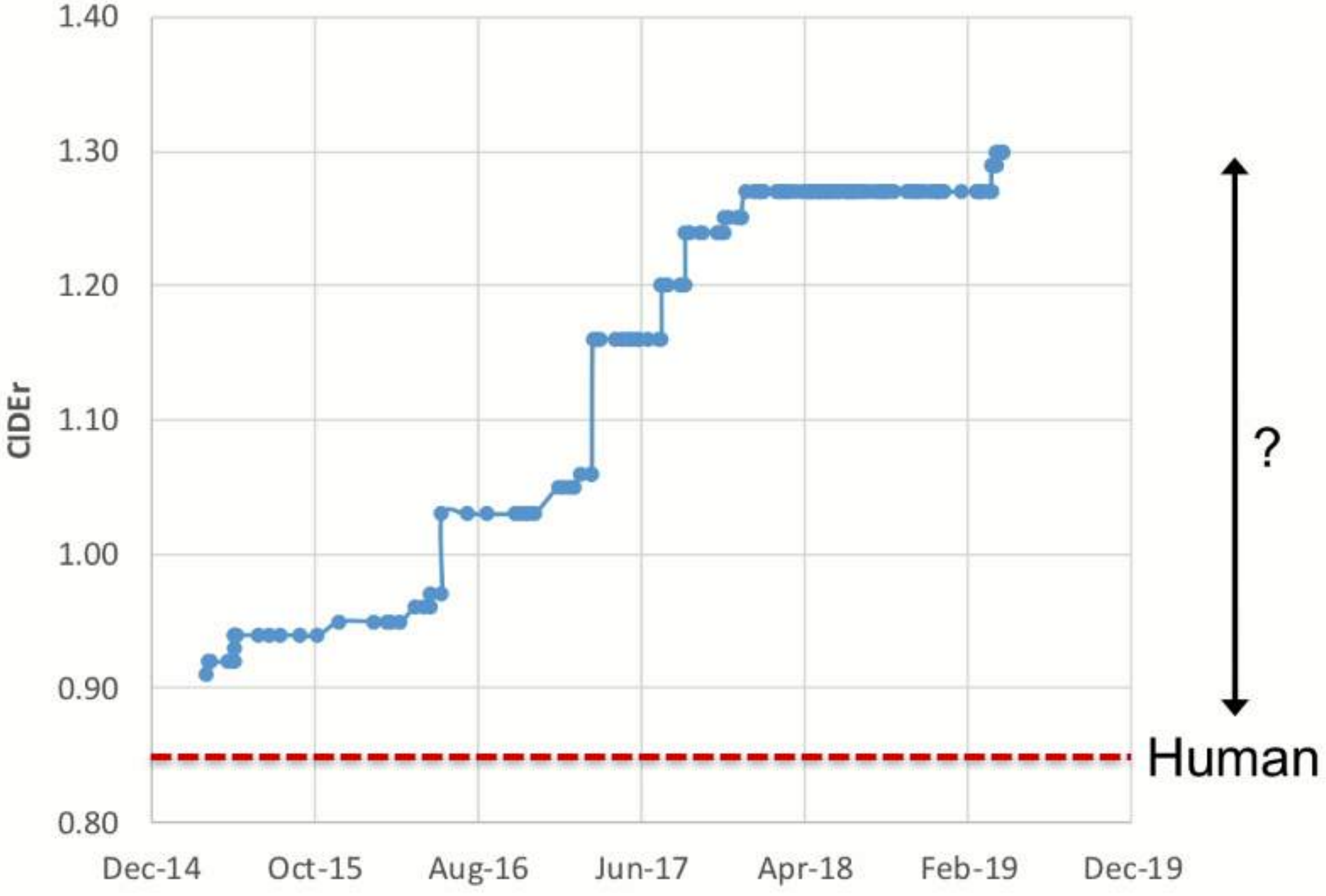
SOTA on COCO Dataset



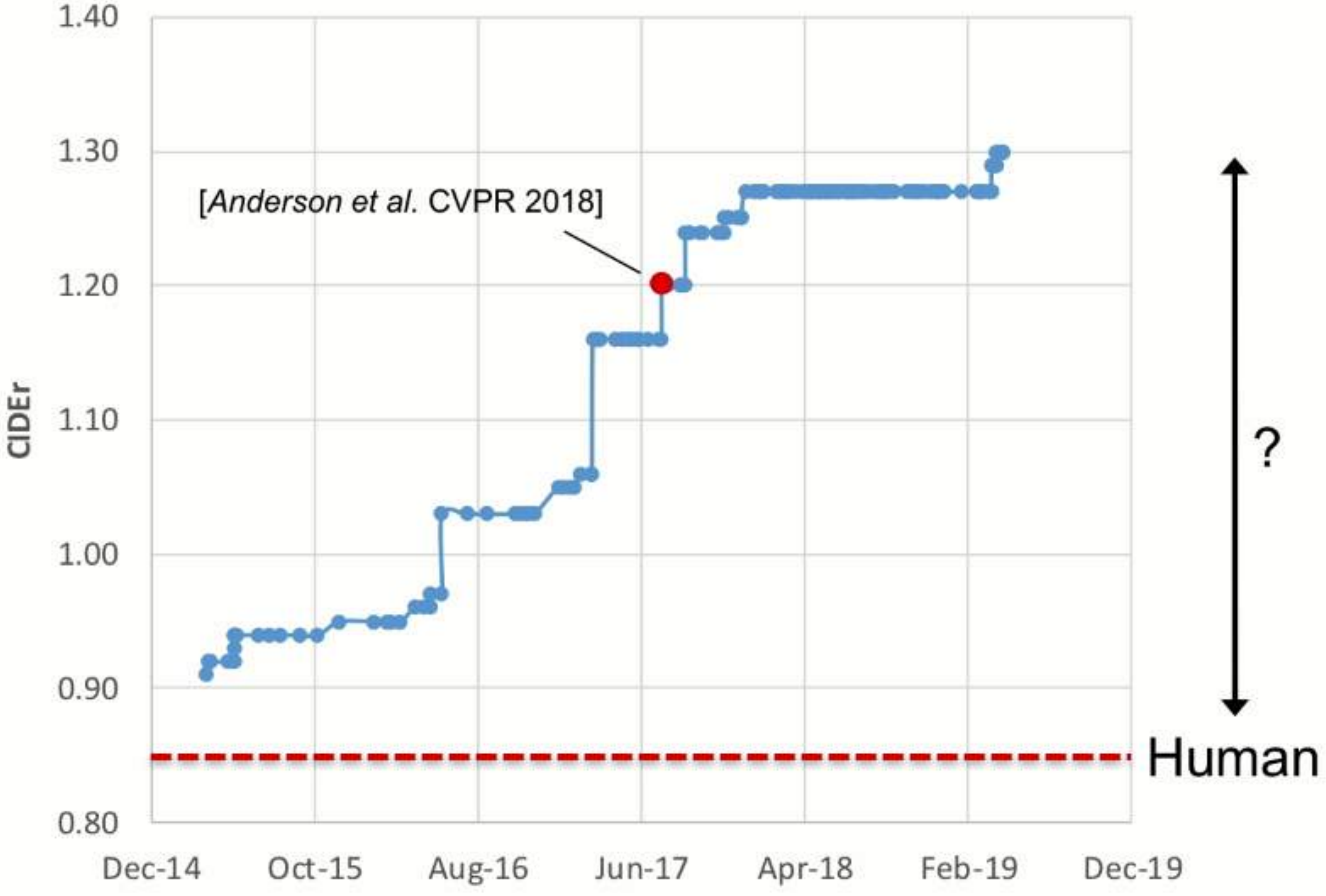
SOTA on COCO Dataset



SOTA on COCO Dataset



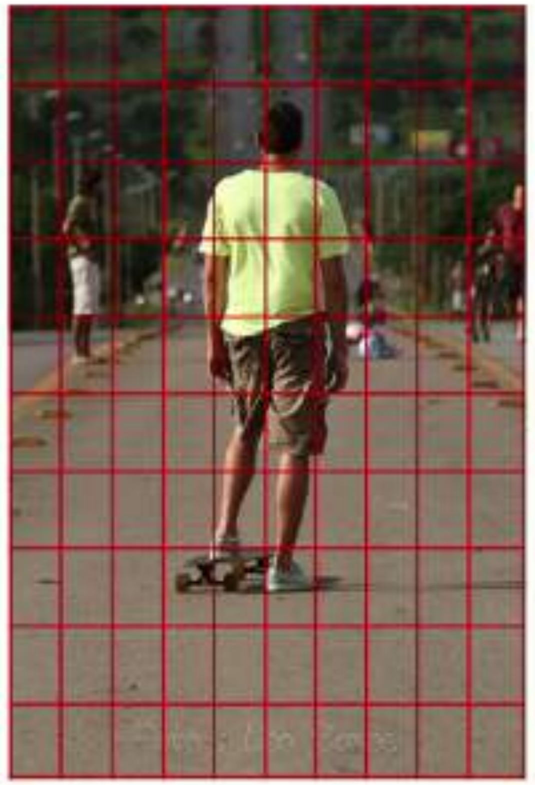
SOTA on COCO Dataset



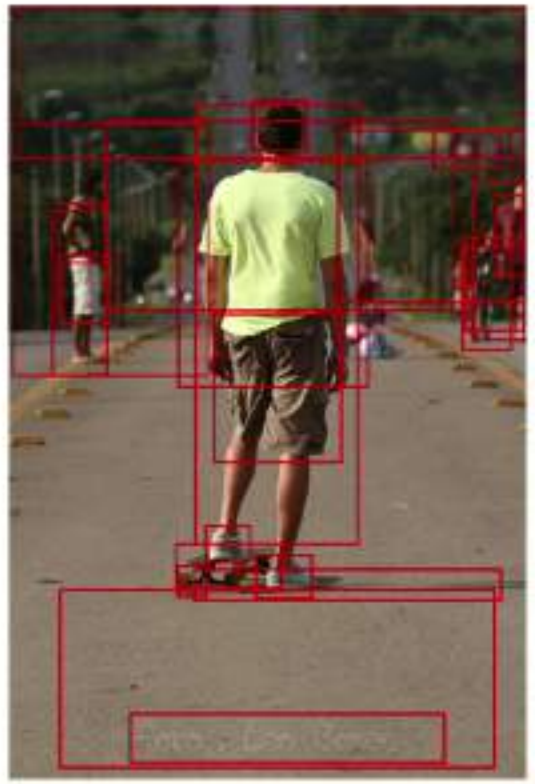
(near) SOTA on COCO Dataset

- Bottom-Up and Top-Down Attention
  - Incorporates object detection into vision & language problems
  - Now the standard approach to image captioning / VQA

10 x 10 regions



Standard attention over spatial output from a CNN



k regions

Up-Down attention over detected objects

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A brown sheep standing in a field of grass.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is playing baseball.

A brown sheep standing in a field of grass.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A zebra is laying down in the grass.

[*Anderson et al.* CVPR 2018]

(near) SOTA on COCO Dataset

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[Anderson et al. CVPR 2018]

Mentions in Training Captions

Two hot dogs on a tray with a drink.

Two elephants and a baby elephant walking together.

A man in a white shirt is ~~playing baseball~~.

A brown sheep standing in a field of grass.

A ~~zebra~~ is laying down in the grass.

[*Anderson et al.* CVPR 2018]

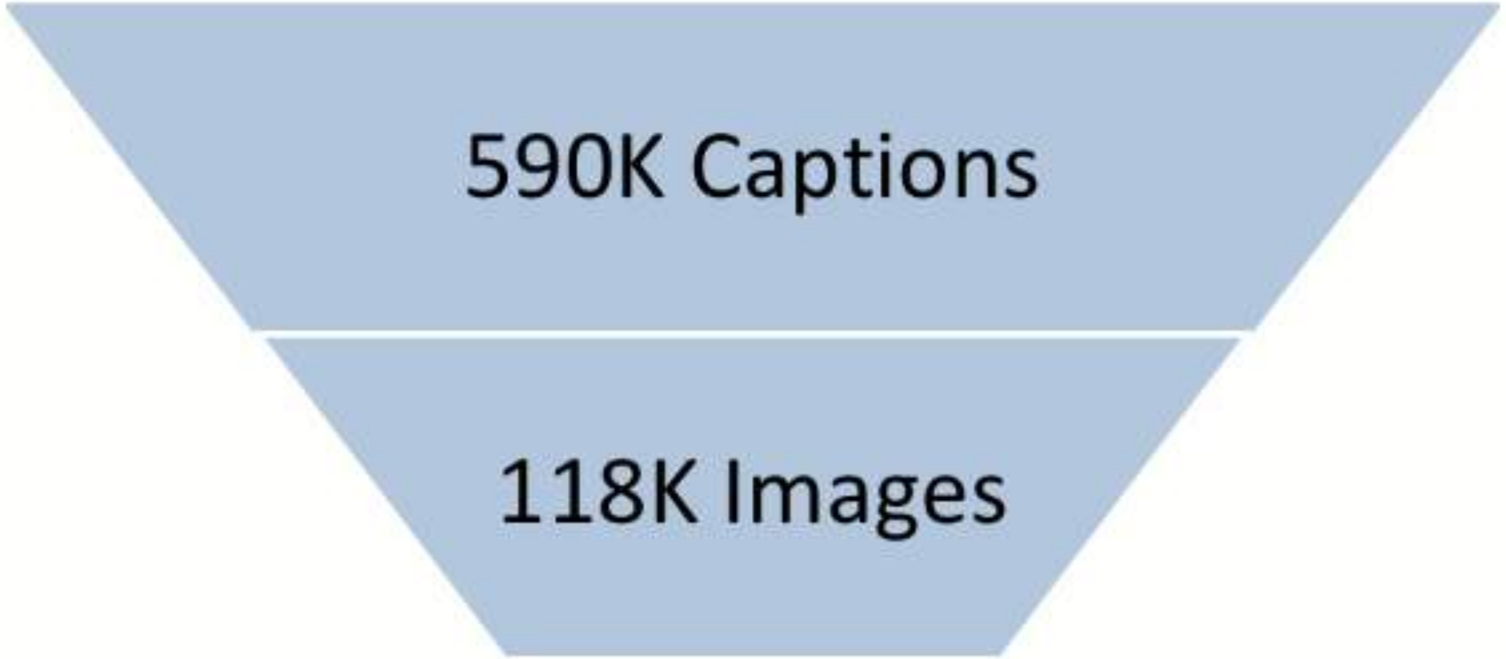
COCO 2017 Training Set

[*Chen et al.* arXiv 1504.00325 2015]

COCO 2017 Training Set

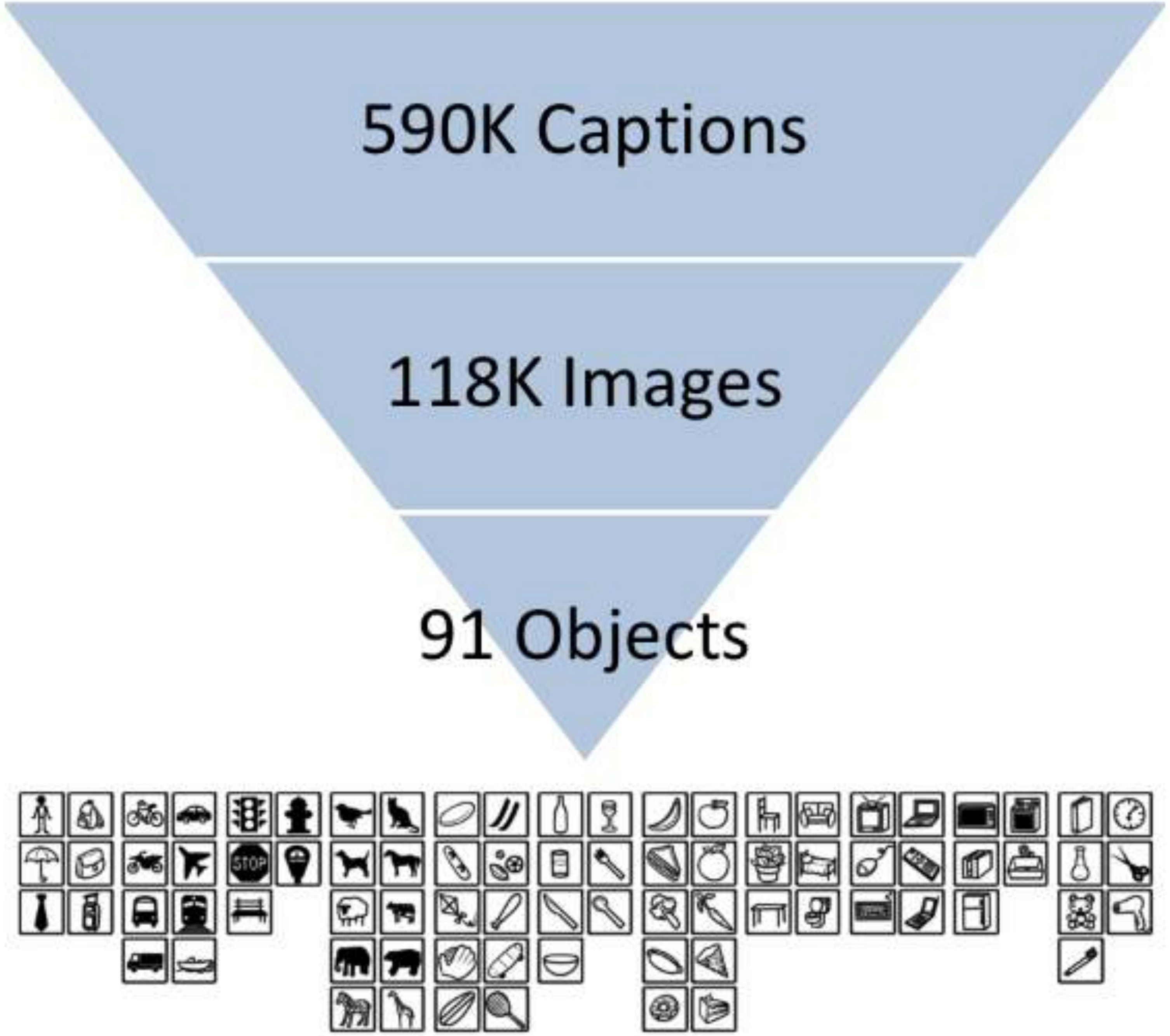
[*Chen et al.* arXiv 1504.00325 2015]

COCO 2017 Training Set



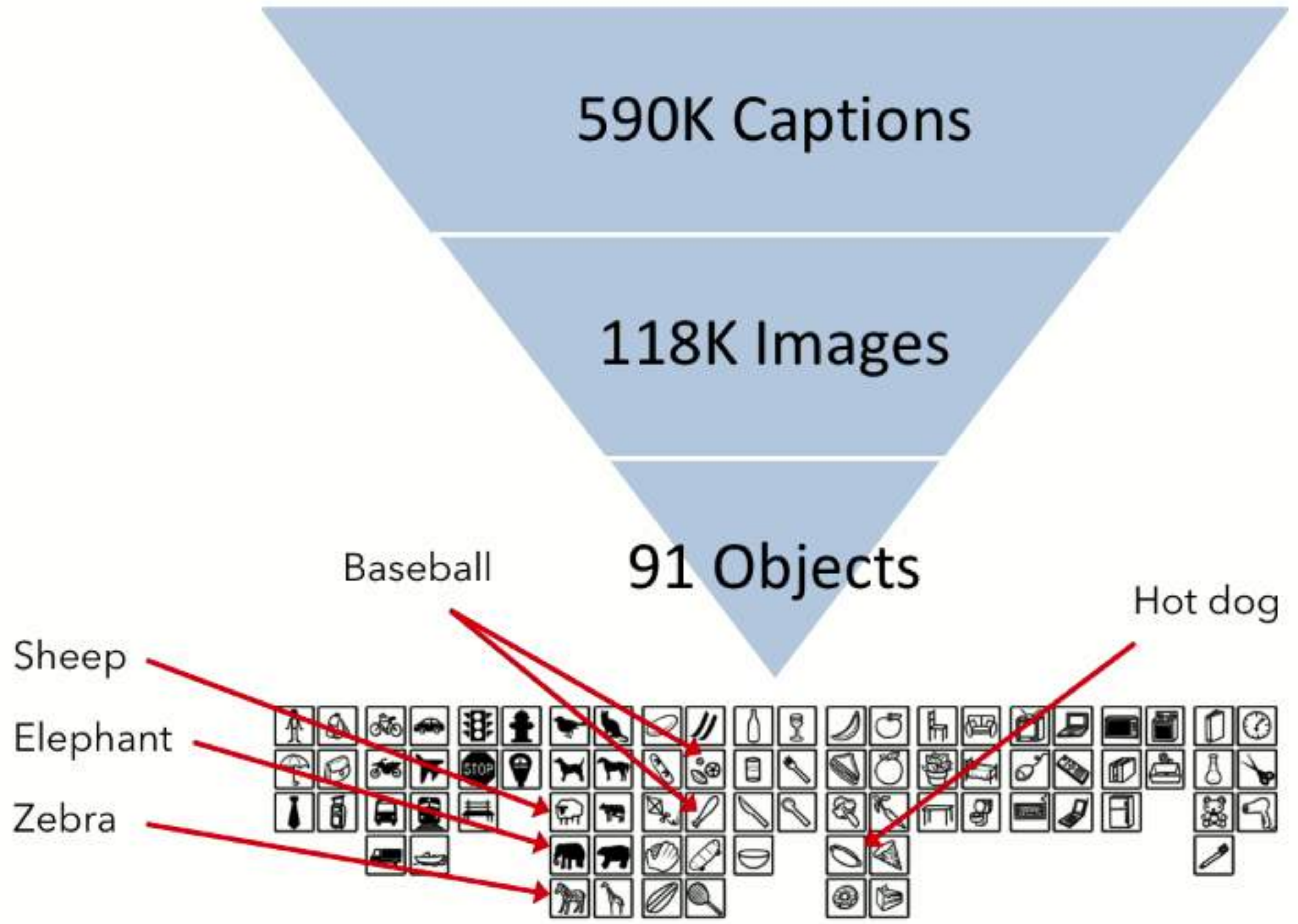
[*Chen et al.* arXiv 1504.00325 2015]

COCO 2017 Training Set



[*Chen et al.* arXiv 1504.00325 2015]

COCO 2017 Training Set

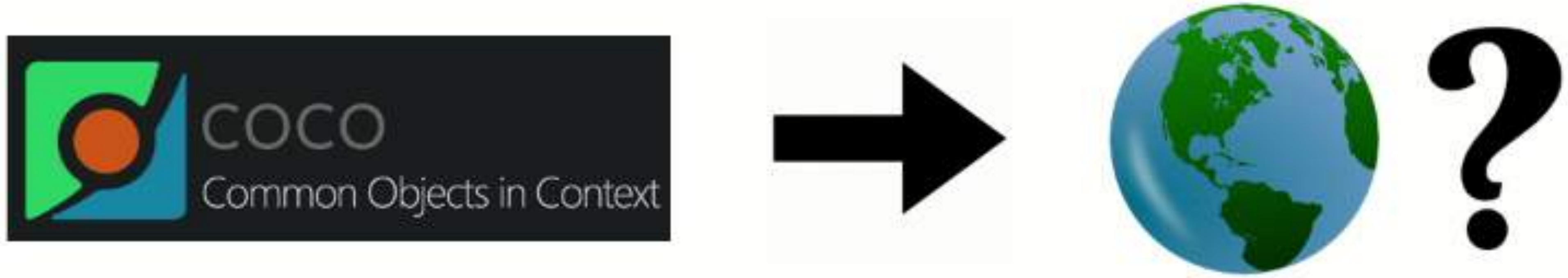


[*Chen et al.* arXiv 1504.00325 2015]

How to scale to images ‘in the wild’?

COCO
Common Objects in Context

How to scale to images ‘in the wild’?



Idea: Collect more data

How to scale to images 'in the wild'?

COCO
Common Objects in Context

?

Idea: Collect more data

How to scale to images 'in the wild'?

COCO Common Objects in Context →  ?

Idea: Collect more data

A man in a white shirt is doing a karate kick.

A tiger is laying down in the grass.

How to scale to images 'in the wild'?

COCO
Common Objects in Context

Idea: Collect more data

- Redundancy / duplicated effort

A man in a white shirt is doing a karate kick.

A tiger is laying down in the grass.

How to scale to images ‘in the wild’?

COCO
Common Objects in Context

Idea: Collect more data

- Redundancy / duplicated effort
- Expense

A man in a white shirt is doing a karate kick.

A tiger is laying down in the grass.

How to scale to images ‘in the wild’?

Alternative: Exploit existing data sources

- Object Detection Datasets
  - Provides missing grounded nouns
  - Orders of magnitude larger than COCO
  - Can be collected semi-automatically, e.g. [*Papadopoulos et al.* ICCV 2017]

[*Kuznetsova et al.* arXiv 1811.00982 2018]

How to scale to images 'in the wild'?

Alternative: Exploit existing data sources

- Object Detection Datasets
  - Provides missing grounded nouns
  - Orders of magnitude larger than COCO
  - Can be collected semi-automatically, e.g. [*Papadopoulos et al.* ICCV 2017]

[*Kuznetsova et al.* arXiv 1811.00982 2018]

- Unaligned Text Corpora
  - Provide (something from DCC paper)
  - Orders of magnitude larger than COCO
  - Can be collected automatically

The tiger has a muscular body with powerful forelimbs, a large head and a tail that is about half the length of its body. Its pelage is dense and heavy, and colouration varies

[Wikipedia]

- Other?

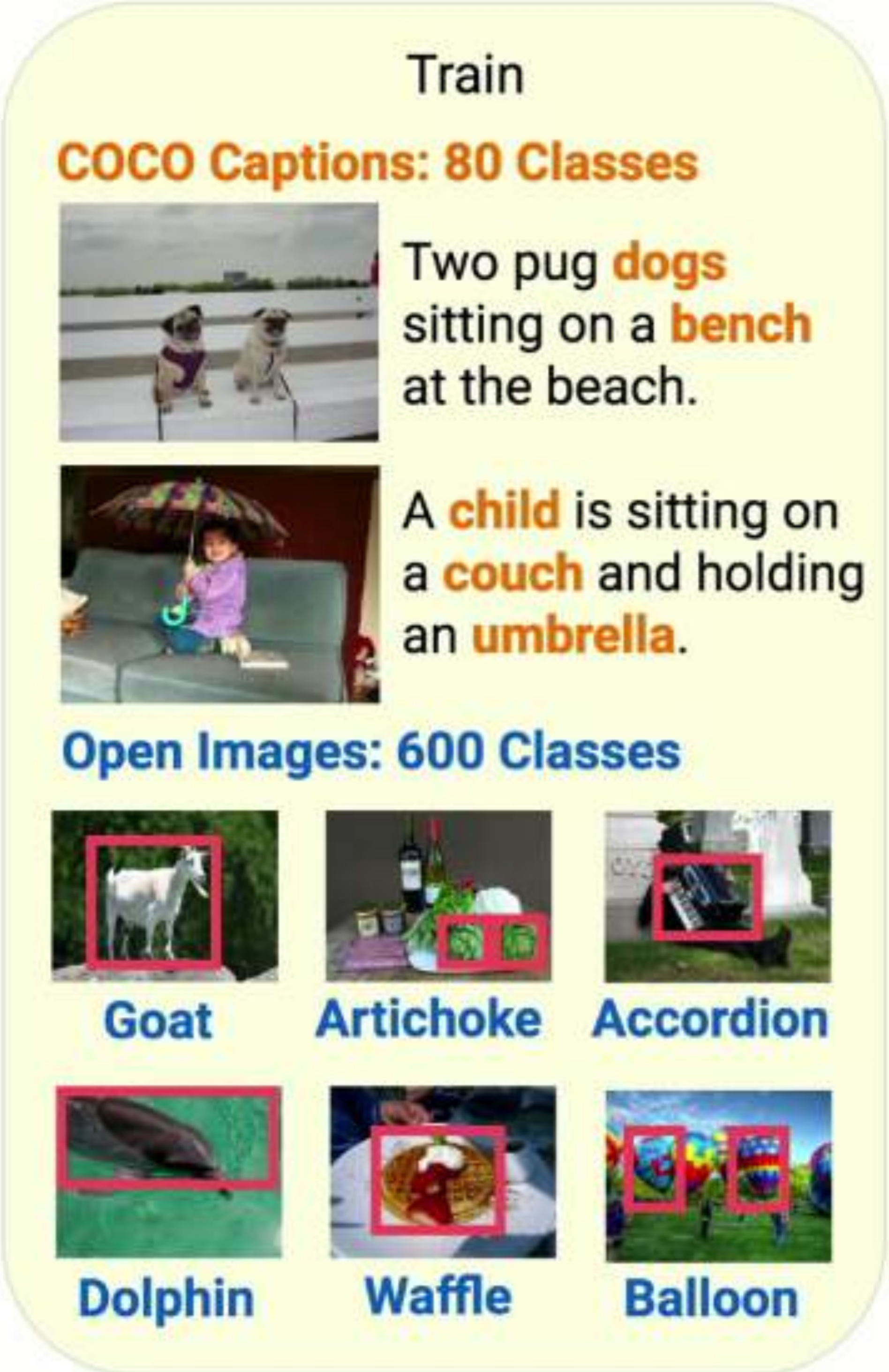
nocaps Benchmark

nocaps.org

nocaps Benchmark

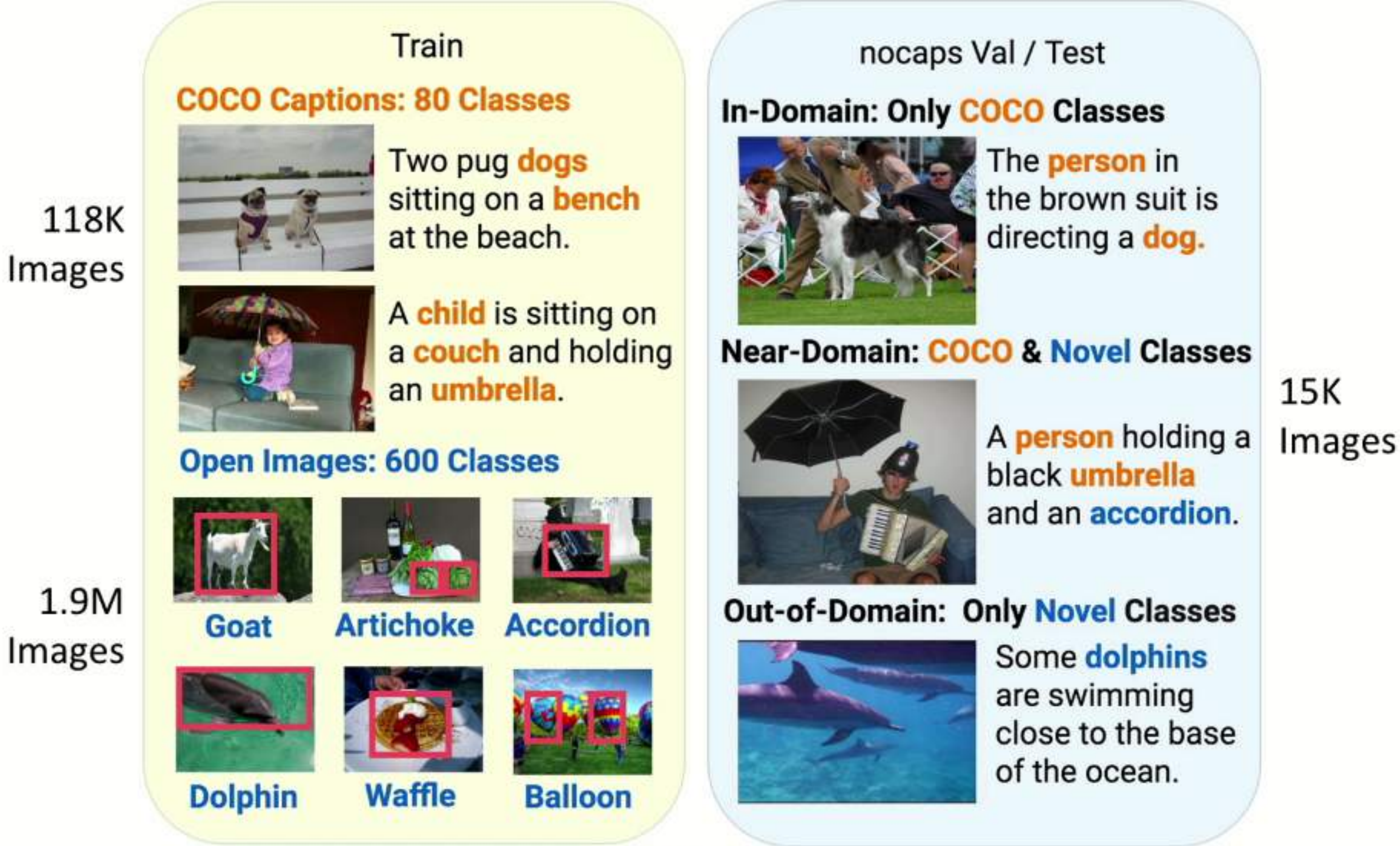
118K Images

1.9M Images



nocaps.org

nocaps Benchmark



nocaps.org

Novel Object Captioning

nocaps.org

Novel Object Captioning

- **nocaps** for **n**ovel **o**bject **cap**tioning at **s**cale

nocaps.org

Novel Object Captioning

- **nocaps** for **n**ovel **o**bject **cap**tioning at **s**cale
- Builds on prior work and community interest in ‘novel object captioning’ [*Hendricks et al.* CVPR 2016]

Novel Object Captioning

- **nocaps** for **n**ovel **o**bject **cap**tioning at **s**cale
- Builds on prior work and community interest in 'novel object captioning' [*Hendricks et al.* CVPR 2016]
- Comparison to existing held-out COCO proof-of-concept dataset:
  - From 8 novel object classes to ~400
  - Novel objects no longer highly similar to known classes (e.g. horse is seen, zebra is novel)
  - From 4K evaluation captions to 151K

nocaps.org

Novel Object Captioning

- **nocaps** for **n**ovel **o**bject **cap**tioning at **s**cale
- Builds on prior work and community interest in ‘novel object captioning’ [*Hendricks et al.* CVPR 2016]
- Comparison to existing held-out COCO proof-of-concept dataset:
  - From 8 novel object classes to ~400
  - Novel objects no longer highly similar to known classes (e.g. horse is seen, zebra is novel)
  - From 4K evaluation captions to 151K
- Learning Perspective: Includes aspects of both domain adaptation and transfer learning.

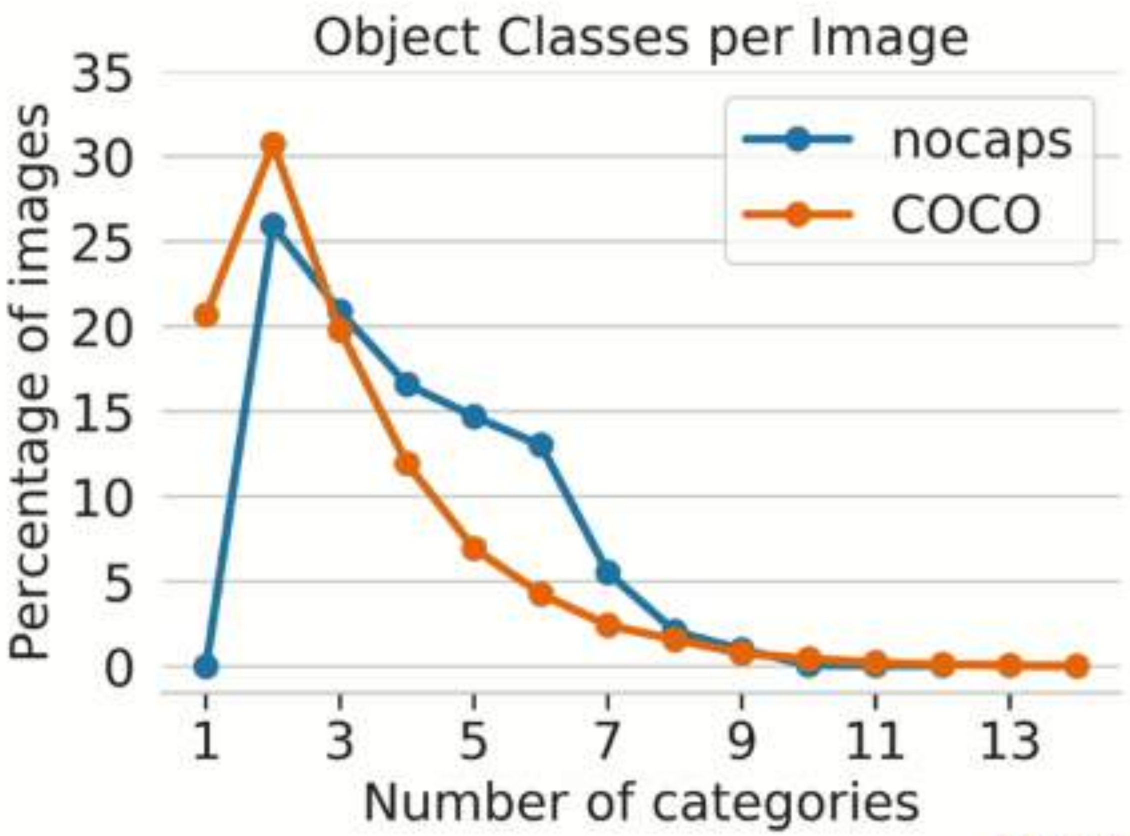
nocaps.org

nocaps Image Selection Strategy

- Select 4.5K/10.6K of 42K/125K Open Images validation/test images for nocaps
  - Prioritizing even class representation and multiple objects per image

nocaps Image Selection Strategy

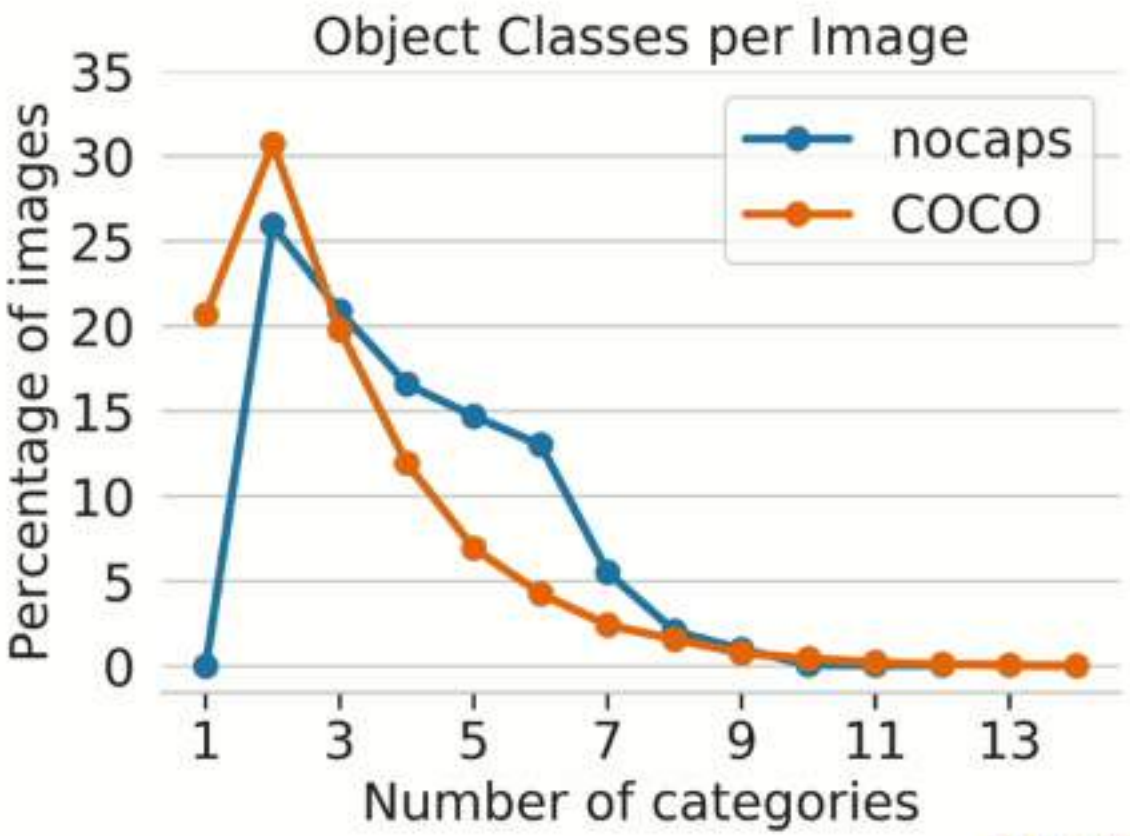
- Select 4.5K/10.6K of 42K/125K Open Images validation/test images for nocaps
  - Prioritizing even class representation and multiple objects per image
- Avg 4.0 object classes per image (COCO 2.9)

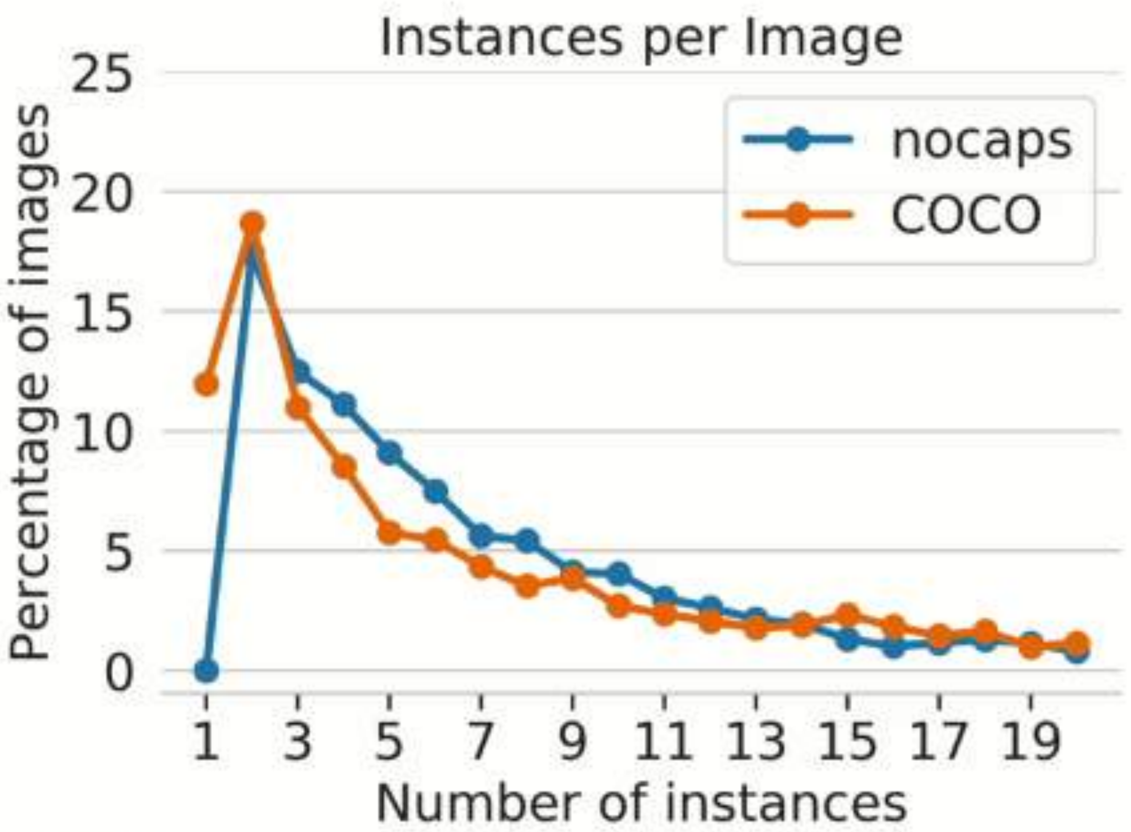


nocaps.org

nocaps Image Selection Strategy

- Select 4.5K/10.6K of 42K/125K Open Images validation/test images for nocaps
  - Prioritizing even class representation and multiple objects per image
- Avg 4.0 object classes per image (COCO 2.9)
- Avg 8.0 object instances per image (COCO 7.4)





nocaps.org

AMT Collection Interface

Describe the image in one sentence

Instructions:

- In each HIT you must describe 5 images.
- Describe all the important parts of the scene.
- The sentence should contain at least 8 words.
- Avoid making spelling errors in your description.
- We provide keywords that may help identify some of the objects in the image.
- It is not mandatory to mention any of the keywords.
- Do not start the sentences with "There is" or "There are".
- Do not write your descriptions as "An image containing...", "A photo of..." or similar.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person in the image might say.
- Do not give people proper names.
- Do not use the text box to report an error with the HIT.

Keywords: cart, person, woman, clothing, building, vegetable

Describe the image in one sentence

Prev (1/5) Next

Shortcuts

Previous: **Alt+K** Next: **Alt+L**

nocaps.org

AMT Collection Interface

Describe the image in one sentence

Instructions:

- In each HIT you must describe 5 images.
- Describe all the important parts of the scene.
- The sentence should contain at least 8 words.
- Avoid making spelling errors in your description.
- We provide keywords that may help identify some of the objects in the image.
- It is not mandatory to mention any of the keywords.
- Do not start the sentences with "There is" or "There are".
- Do not write your descriptions as "An image containing...", "A photo of..." or similar.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person in the image might say.
- Do not give people proper names.
- Do not use the text box to report an error with the HIT.

Keywords: cart, person, woman, clothing, building, vegetable

Describe the image in one sentence

Workers primed with correct object classes

Shortcuts

Previous: **Alt+K**

(1/5)

Next

nocaps.org

AMT Collection Interface

Describe the image in one sentence

Instructions:

- In each HIT you must describe 5 images.
- Describe all the important parts of the scene.
- The sentence should contain at least 8 words.
- Avoid making spelling errors in your description.
- We provide keywords that may help identify some of the objects in the image.
- It is not mandatory to mention any of the keywords.
- Do not start the sentences with "There is" or "There are".
- Do not write your descriptions as "An image containing...", "A photo of..." or similar.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person in the image might say.
- Do not give people proper names.
- Do not use the text box to report an error with the HIT.

Keywords: cart, person, woman, clothing, building, vegetable

Describe the image in one sentence

Workers primed with correct object classes

Shortcuts

Previous: **Alt+K**

(1/5)

Next

nocaps.org

Impact of Priming

- Higher quality reference captions

nocaps.org

Impact of Priming

- Higher quality reference captions

Object classes: *Red Panda, Tree*

No Priming: *A brown rodent climbing up a tree in the woods.*

With Priming: *A red panda is sitting in grass next to a tree.*

nocaps.org

Impact of Priming

- Higher quality reference captions

Object classes: *Red Panda, Tree*

No Priming: *A brown rodent climbing up a tree in the woods.*

With Priming: *A red panda is sitting in grass next to a tree.*

Unique n-grams in equally-sized dataset samples:

Dataset	1-grams	2-grams	3-grams	4-grams
COCO	6,913	46,664	92,946	119,582
nocaps	8,291	59,714	116,765	144,577

nocaps.org

AMT Collection Interface

Describe the image in one sentence

Instructions:

- In each HIT you must describe 5 images.
- Describe all the important parts of the scene.
- The sentence should contain at least 8 words.
- Avoid making spelling errors in your description.
- We provide keywords that may help identify some of the objects in the image.
- It is not mandatory to mention any of the keywords.
- Do not start the sentences with "There is" or "There are".
- Do not write your descriptions as "An image containing...", "A photo of..." or similar.
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person in the image might say.
- Do not give people proper names.
- Do not use the text box to report an error with the HIT.

Keywords: cart, person, woman, clothing, building, vegetable

Describe the image in one sentence

Workers primed with correct object classes

Shortcuts

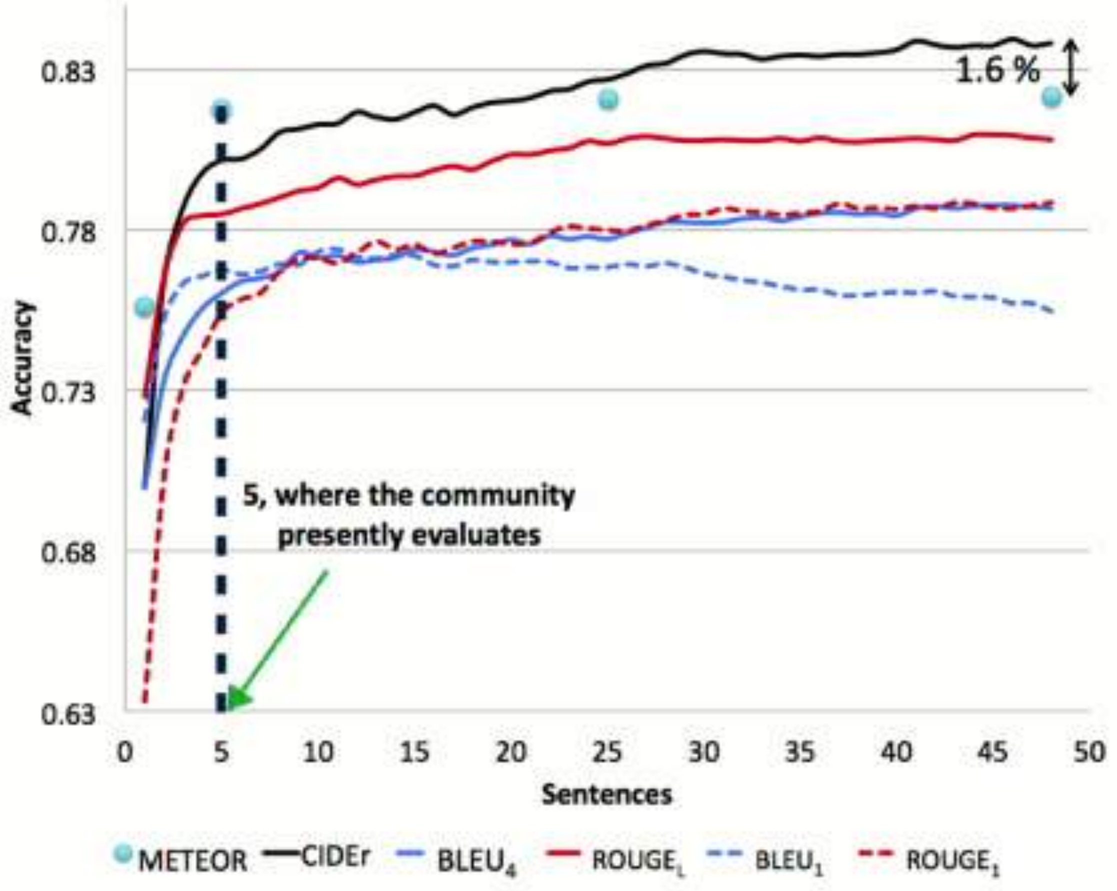
Previous: **Alt+K**

(1/5)

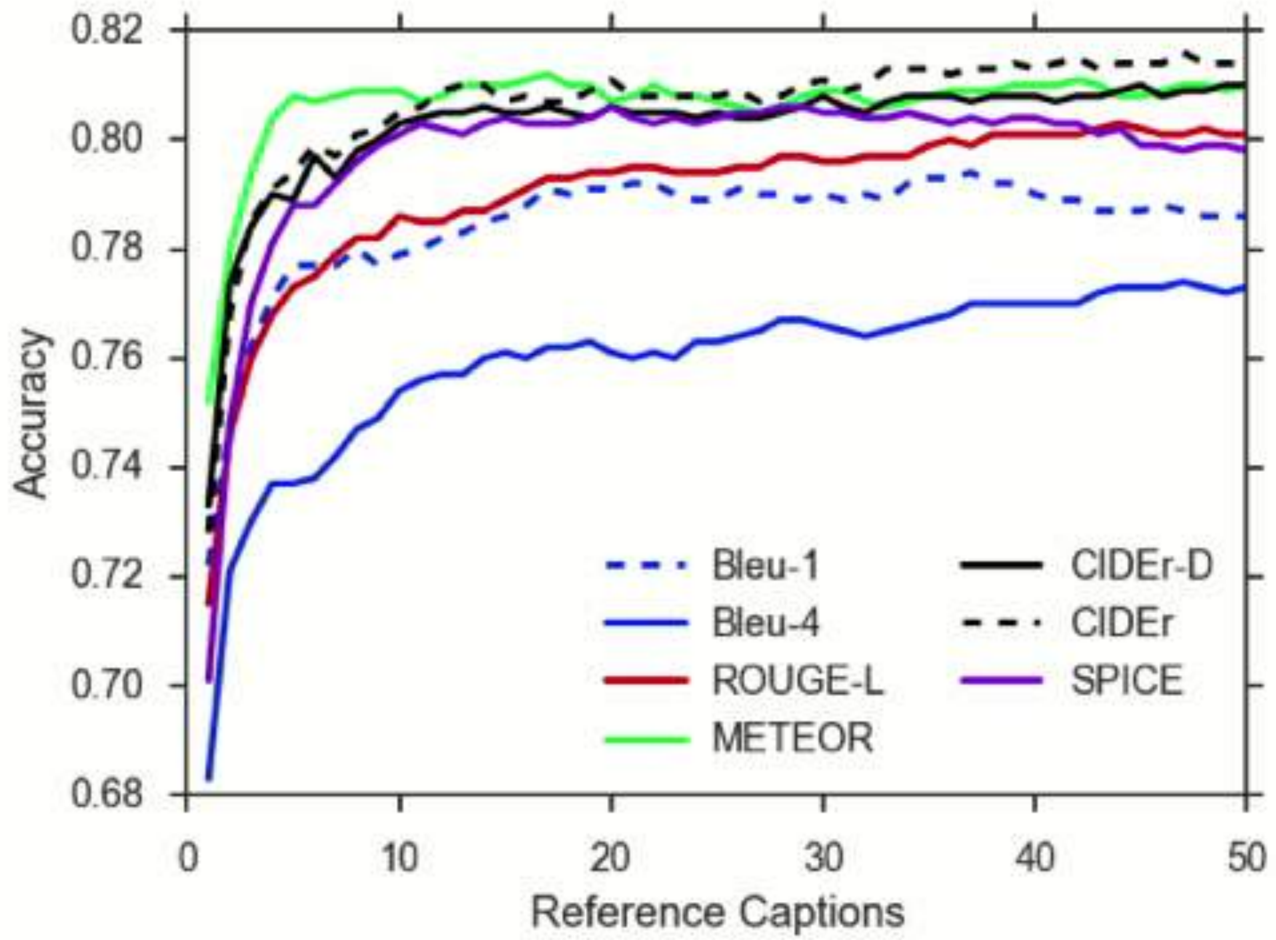
Next

nocaps.org

How Many Reference Captions?



[*Vedantam et al.* CVPR 2015]

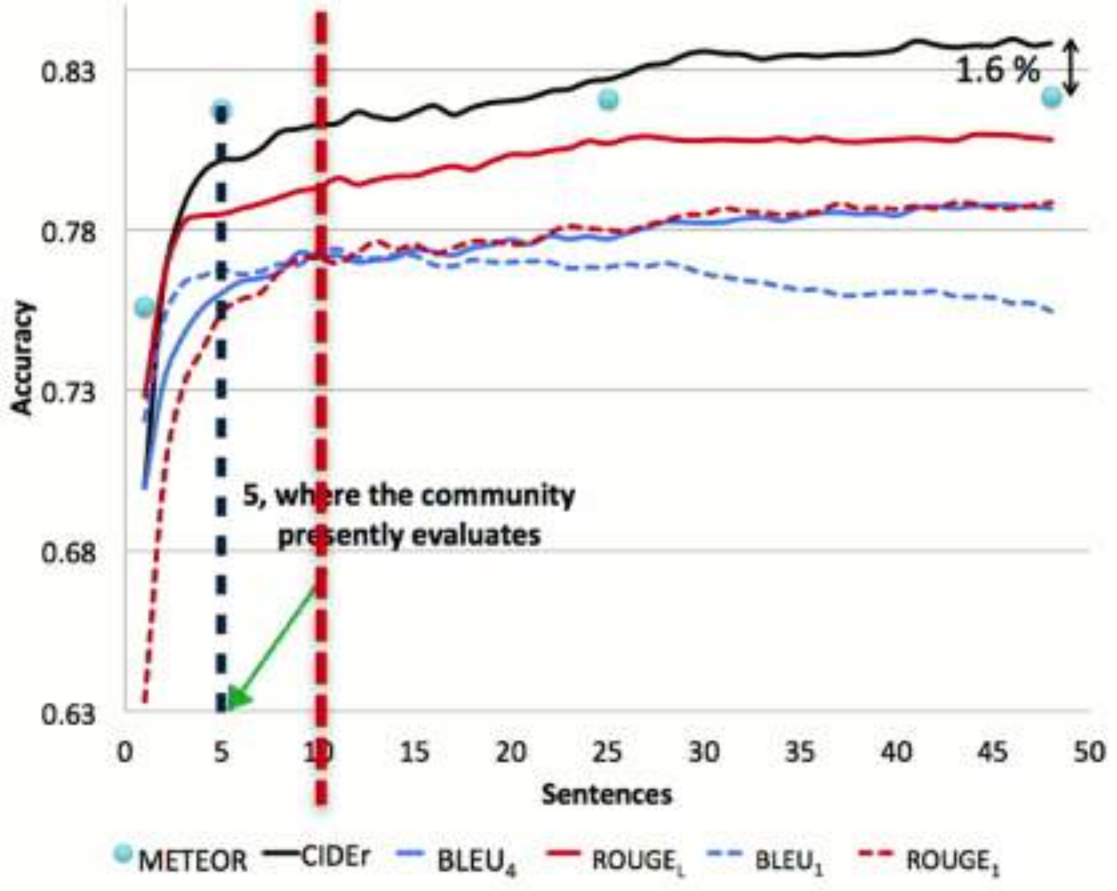


[*Anderson et al.* ECCV 2016]

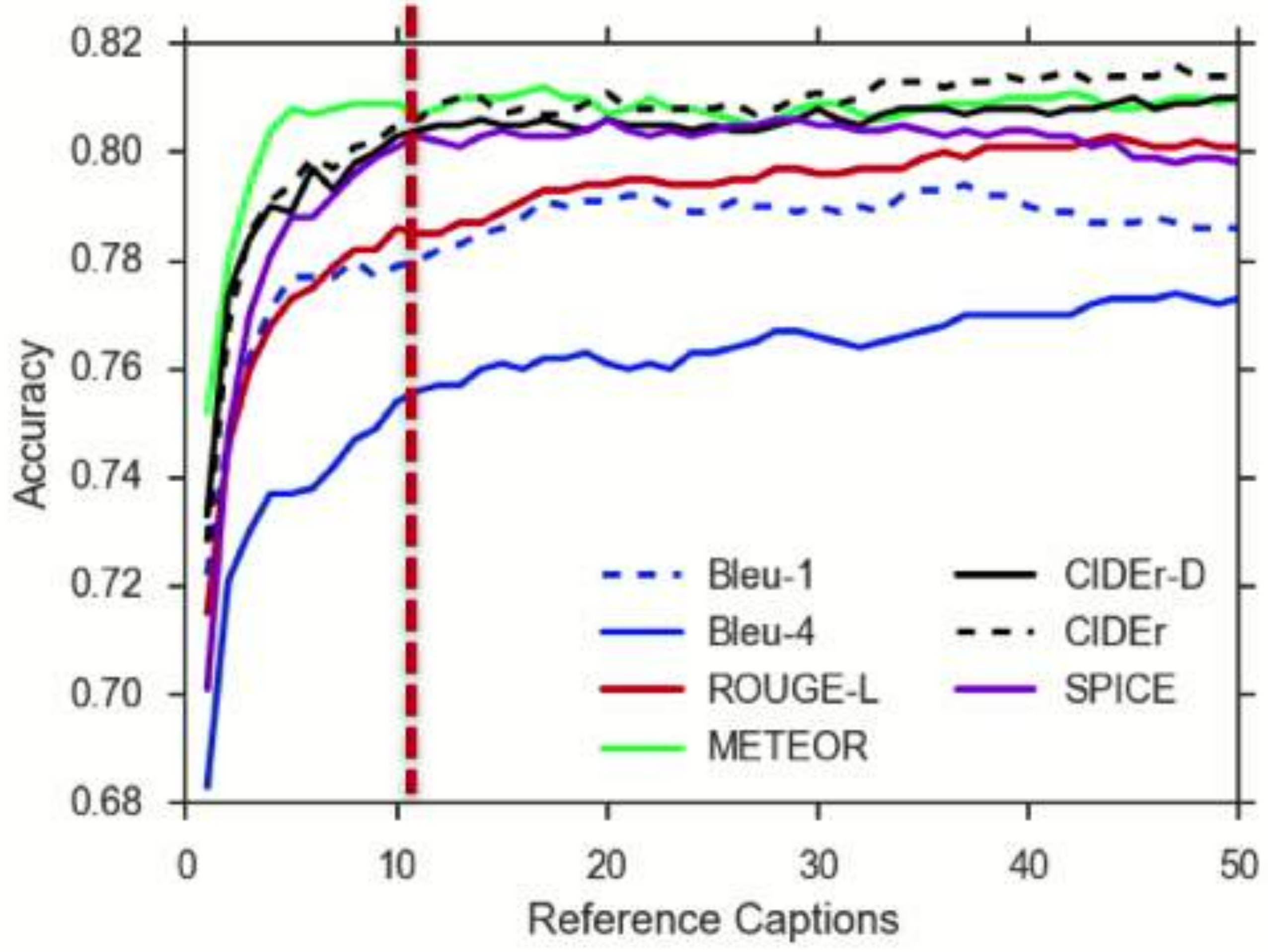
nocaps.org

How Many Reference Captions?

- 10 reference captions appears to be the sweet spot



[*Vedantam et al.* CVPR 2015]

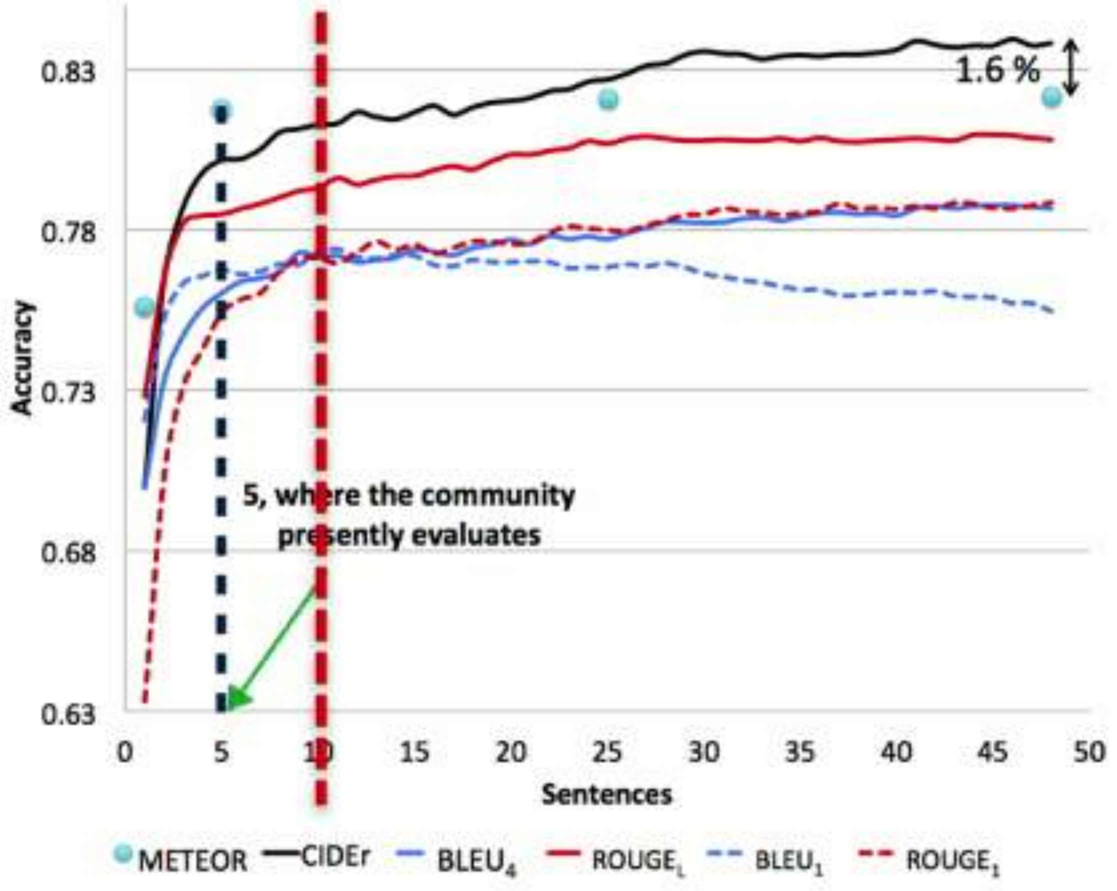


[*Anderson et al.* ECCV 2016]

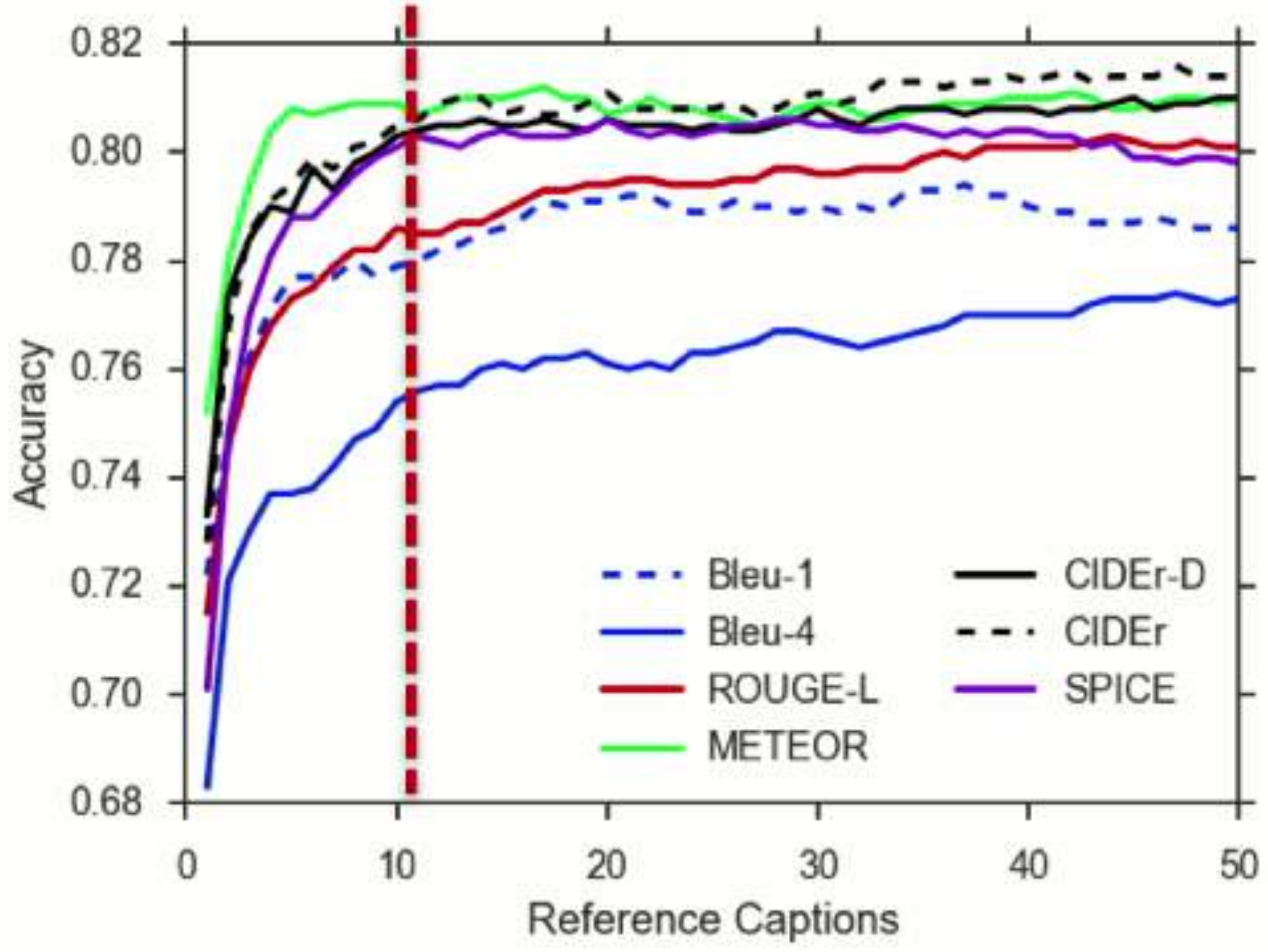
nocaps.org

How Many Reference Captions?

- 10 reference captions appears to be the sweet spot
- Collect 166K captions (10 per image plus a human baseline)



[*Vedantam et al.* CVPR 2015]



[*Anderson et al.* ECCV 2016]

nocaps.org

Impact of Priming

- Higher quality reference captions

Object classes: *Red Panda, Tree*

No Priming: *A brown rodent climbing up a tree in the woods.*

With Priming: *A red panda is sitting in grass next to a tree.*

nocaps.org

nocaps Example: near-domain

Color Key:

COCO object class

Open Images object class

1. *A **man** sitting in the saddle on a **camel**.*
2. *A **person** is sitting on a **camel** with another **camel** behind him.*
3. *A **man** with long hair and blue jeans sitting on a **camel**.*
4. ***Man** sitting on a **camel** with a standing **camel** behind them.*
5. *Long haired **man** wearing sitting on blanket draped **camel.***
6. *A camel stands behind a sitting **camel** with a **man** on its back.*
7. *The standing **camel** is near a sitting one with a **man** on its back.*
8. *Someone is sitting on a camel and is in front of another **camel**.*
9. *Two **camels** in the dessert and a **man** sitting on the sitting one.*
10. *Two **camels** are featured in the sand with a **man** sitting on one of the seated **camels**.*

nocaps Example: out-of-domain

Color Key:

COCO object class

Open Images object class

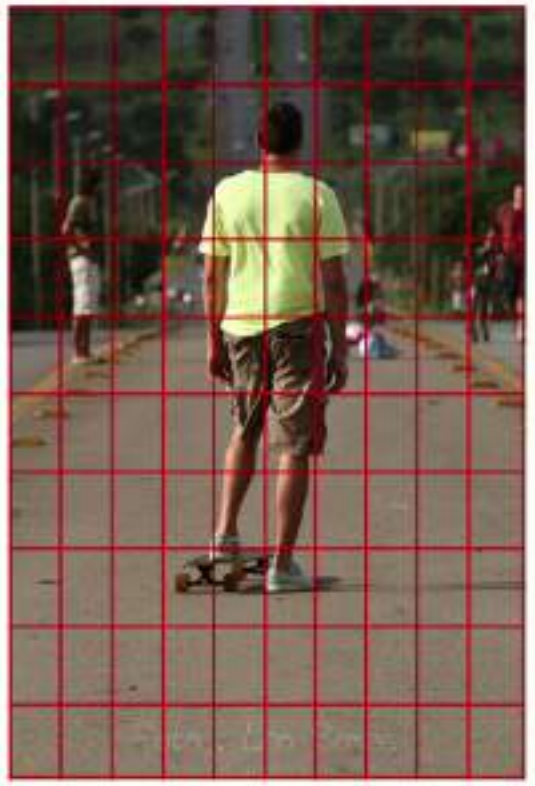
1. *A **tank** vehicle stopped at a gas station.*
2. *A **tank** and a military jeep at a gas station*
3. *A jeep and a tan colored **tank** getting gas at a gas station.*
4. *A **tank** and a **truck** sit at a gas station pump.*
5. *An Army humvee is at getting gas from the 76 gas station.*
6. *An army **tank** is parked at a gas station.*
7. *A **land vehicle** is parked in a gas station fueling.*
8. *A large military vehicle at the gas pump of a gas station.*
9. *A tanker parked outside of an old gas station*
10. *Multiple military vehicles getting gasoline at a civilian gas station.*

nocaps.org

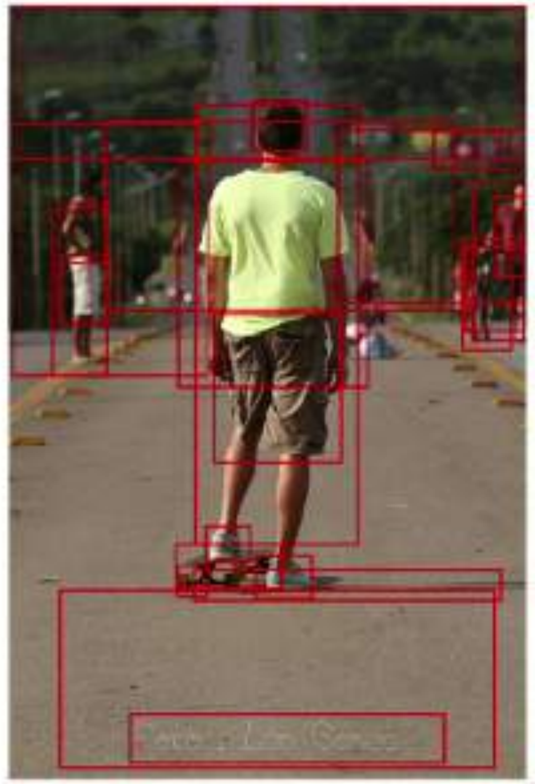
Baseline model (Up-Down)

- Captioning model using Bottom-Up and Top-Down Attention
  - trained only on COCO

10 x 10 regions



Standard attention over spatial output from a CNN



k regions

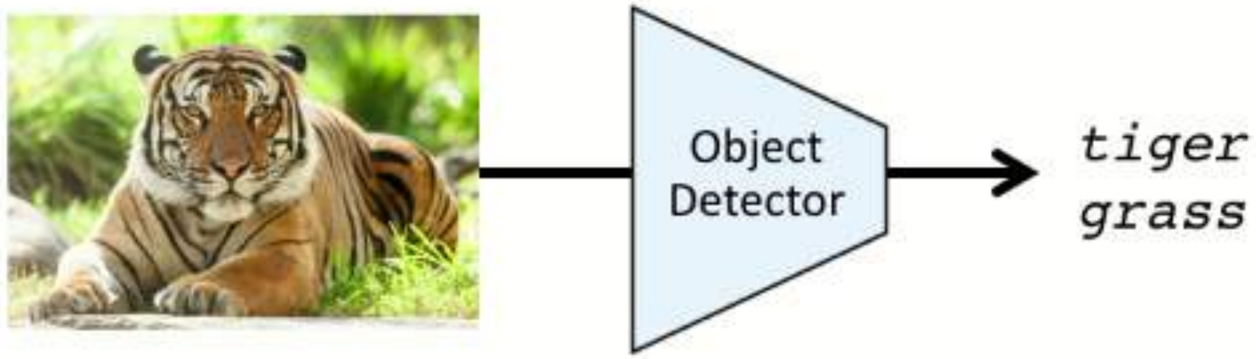
Up-Down attention over detected objects

[*Anderson et al.* CVPR 2018]

nocaps.org

Constrained Beam Search (CBS)

- Key idea: Combine the captioning model with an object detector trained on novel objects

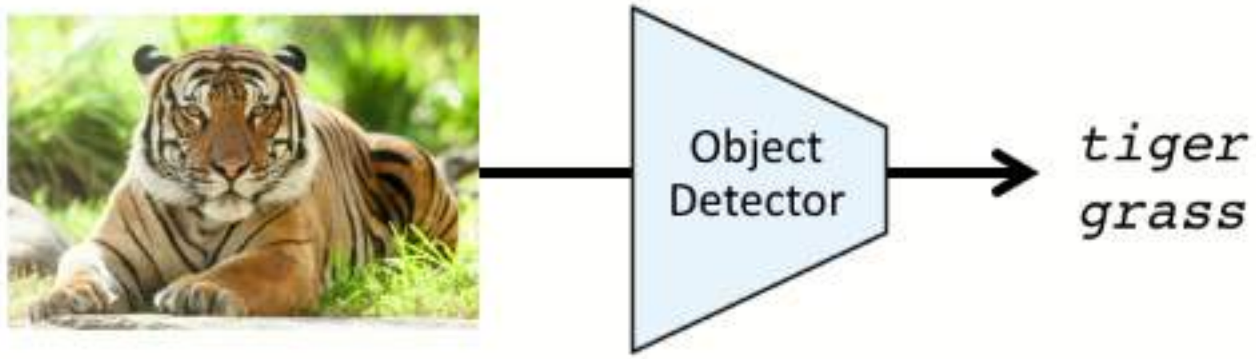


[*Anderson et al.* EMNLP 2017]

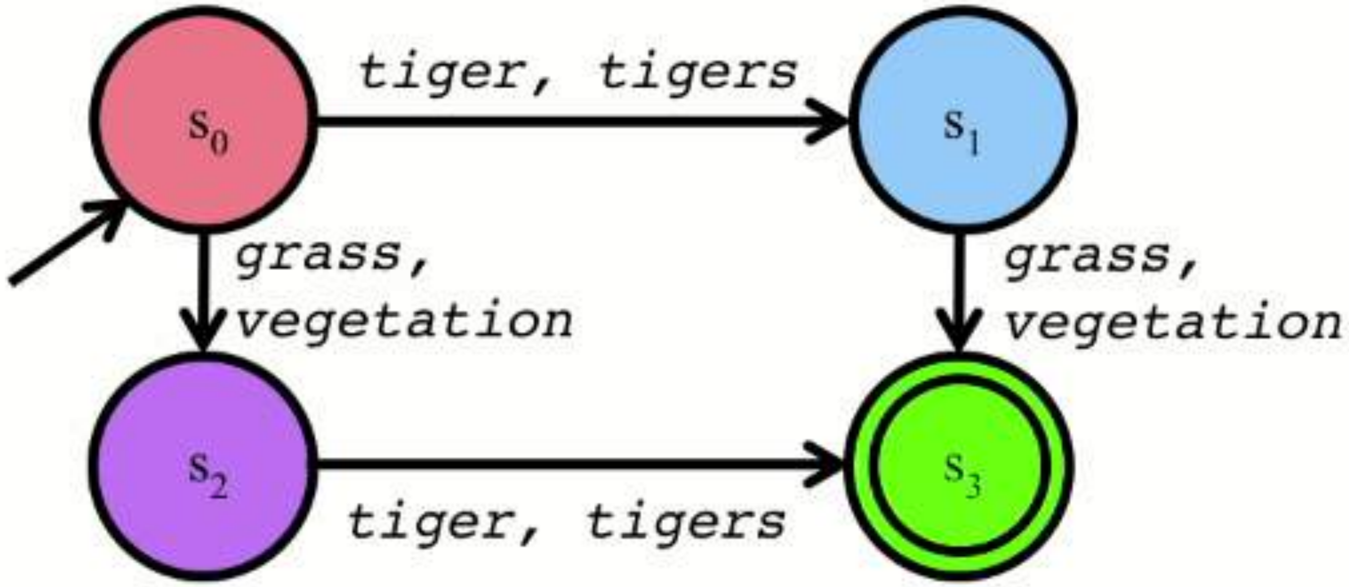
nocaps.org

Constrained Beam Search (CBS)

- Key idea: Combine the captioning model with an object detector trained on novel objects



- Encode the detected objects (plus plurals, synonyms etc.) in a Finite State Machine (FSM)



[*Anderson et al.* EMNLP 2017]

nocaps.org

Constrained Beam Search (CBS)

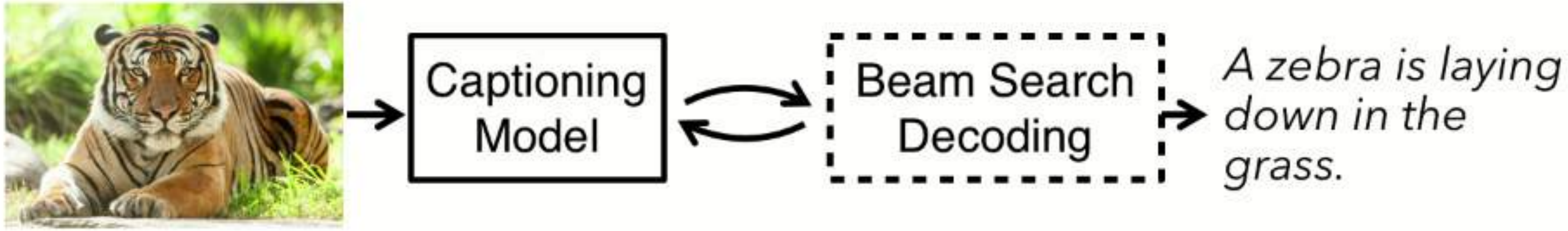
- Beam Search finds high probability output sequences

[*Anderson et al.* EMNLP 2017]

nocaps.org

Constrained Beam Search (CBS)

- Beam Search finds high probability output sequences
- CBS finds finds high probability sequences that satisfy the (hard) constraints given by the FSM

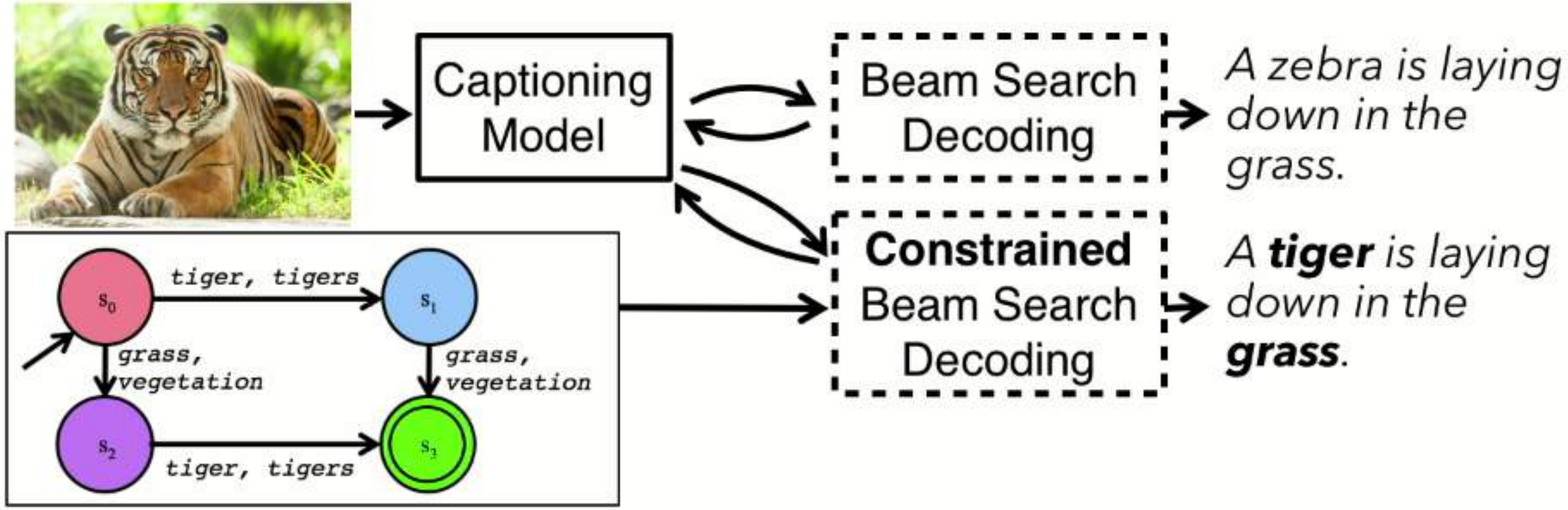


[*Anderson et al.* EMNLP 2017]

nocaps.org

Constrained Beam Search (CBS)

- Beam Search finds high probability output sequences
- CBS finds finds high probability sequences that satisfy the (hard) constraints given by the FSM

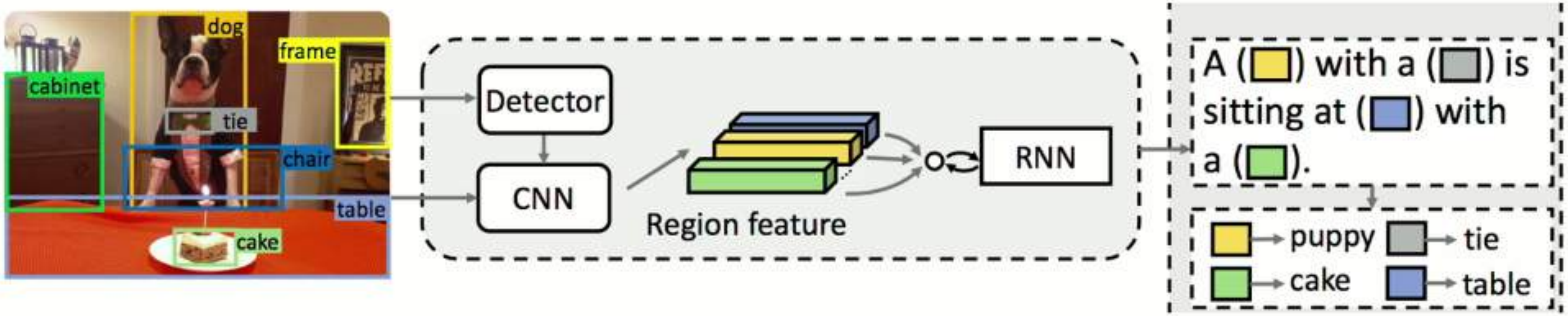


[*Anderson et al.* EMNLP 2017]

nocaps.org

Neural Baby Talk (NBT)

- Generates a sentence template with slots grounded to image regions, then fills slots using an object detector.

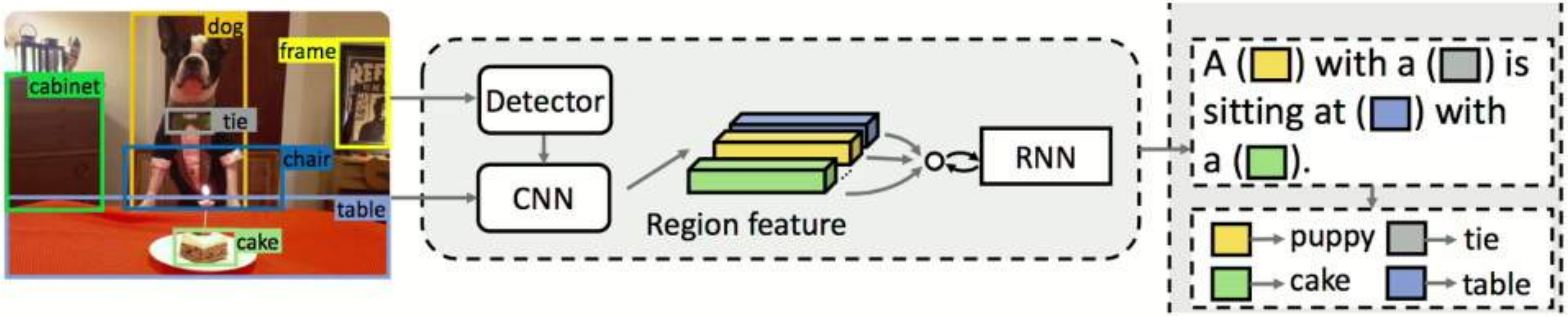


[*Lu et al.* CVPR 2018]

nocaps.org

Neural Baby Talk (NBT)

- Generates a sentence template with slots grounded to image regions, then fills slots using an object detector.
- Use of an explicit object detector makes the model applicable to novel object captioning



[*Lu et al.* CVPR 2018]

nocaps.org

Results

Method	COCO val 2017					nocaps val							
			Overall			In-Domain		Near-Domain		Out-of-Domain		Overall	
	Bleu-1	Bleu-4	Meteor	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
(1) Up-Down	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	77.6	11.6	58.4	10.4	32.3	8.3	55.8	10.2
(2) Up-Down + CBS	73.3	32.4	25.8	97.7	18.7	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
(3) Up-Down + ELMo + CBS	72.4	31.5	25.7	95.4	18.2	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
(4) Up-Down + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.2	31.5	25.3	95.1	18.0	62.6	10.0	52.7	9.4	51.8	8.6	54.0	9.3
(6) NBT + CBS	70.2	28.2	25.1	92.8	18.1	62.1	10.1	58.3	9.4	62.4	8.9	60.2	9.5
(7) NBT + CBS + GT	-	-	-	-	-	62.4	10.1	59.7	9.5	64.9	9.1	62.3	9.6
(8) Human	66.3	21.7	25.2	85.4	19.8	**84.4**	**14.3**	**85.0**	**14.3**	**95.7**	**14.0**	**87.1**	**14.2**

nocaps.org

Results

Method	COCO val 2017					nocaps val							
			Overall			In-Domain		Near-Domain		Out-of-Domain		Overall	
	Bleu-1	Bleu-4	Meteor	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
(1) Up-Down	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	77.6	11.6	58.4	10.4	32.3	8.3	55.8	10.2
(2) Up-Down + CBS	73.3	32.4	25.8	97.7	18.7	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
(3) Up-Down + ELMo + CBS	72.4	31.5	25.7	95.4	18.2	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
(4) Up-Down + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.2	31.5	25.3	95.1	18.0	62.6	10.0	52.7	9.4	51.8	8.6	54.0	9.3
(6) NBT + CBS	70.2	28.2	25.1	92.8	18.1	62.1	10.1	58.3	9.4	62.4	8.9	60.2	9.5
(7) NBT + CBS + GT	-	-	-	-	-	62.4	10.1	59.7	9.5	64.9	9.1	62.3	9.6
(8) Human	66.3	21.7	25.2	85.4	19.8	**84.4**	**14.3**	**85.0**	**14.3**	**95.7**	**14.0**	**87.1**	**14.2**

- **How hard is the task?** Our best model improves substantially over a COCO-trained baseline, but is still well behind human performance.

nocaps.org

Results

Method	COCO val 2017					nocaps val							
	Overall					In-Domain		Near-Domain		Out-of-Domain		Overall	
	Bleu-1	Bleu-4	Meteor	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
(1) Up-Down	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	77.6	11.6	58.4	10.4	32.3	8.3	55.8	10.2
(2) Up-Down + CBS	73.3	32.4	25.8	97.7	18.7	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
(3) Up-Down + ELMo + CBS	72.4	31.5	25.7	95.4	18.2	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
(4) Up-Down + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.2	31.5	25.3	95.1	18.0	62.6	10.0	52.7	9.4	51.8	8.6	54.0	9.3
(6) NBT + CBS	70.2	28.2	25.1	92.8	18.1	62.1	10.1	58.3	9.4	62.4	8.9	60.2	9.5
(7) NBT + CBS + GT	-	-	-	-	-	62.4	10.1	59.7	9.5	64.9	9.1	62.3	9.6
(8) Human	66.3	21.7	25.2	85.4	19.8	**84.4**	**14.3**	**85.0**	**14.3**	**95.7**	**14.0**	**87.1**	**14.2**

- **Do better language models help?** Incorporating strong language models (e.g. ELMo) helps, mainly on the Out-of-Domain subset.

nocaps.org

Results

Method	COCO val 2017					nocaps val							
	Overall					In-Domain		Near-Domain		Out-of-Domain		Overall	
	Bleu-1	Bleu-4	Meteor	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
(1) Up-Down	**77.0**	**37.2**	**27.8**	**116.2**	**21.0**	77.6	11.6	58.4	10.4	32.3	8.3	55.8	10.2
(2) Up-Down + CBS	73.3	32.4	25.8	97.7	18.7	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
(3) Up-Down + ELMo + CBS	72.4	31.5	25.7	95.4	18.2	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
(4) Up-Down + ELMo + CBS + GT	-	-	-	-	-	84.2	12.6	82.1	11.9	86.7	10.6	83.3	11.8
(5) NBT	72.2	31.5	25.3	95.1	18.0	62.6	10.0	52.7	9.4	51.8	8.6	54.0	9.3
(6) NBT + CBS	70.2	28.2	25.1	92.8	18.1	62.1	10.1	58.3	9.4	62.4	8.9	60.2	9.5
(7) NBT + CBS + GT	-	-	-	-	-	62.4	10.1	59.7	9.5	64.9	9.1	62.3	9.6
(8) Human	66.3	21.7	25.2	85.4	19.8	**84.4**	**14.3**	**85.0**	**14.3**	**95.7**	**14.0**	**87.1**	**14.2**

- **Do better object detectors help?** Supplying ground-truth object detections to the models (in place of Faster R-CNN detections) produces large gains.

nocaps.org

Examples

Up-Down	*A man in a red shirt holding a baseball bat.*	*A bird on the ocean in the ocean.*
Up-Down + ELMo + CBS	*A man in a red hat holding a baseball rifle.*	*A dolphin swimming in the ocean on a sunny day.*
NBT + CBS	*A baseball player holding a baseball rifle in the field.*	*A marine mammal sitting on a dolphin in the ocean.*
Human	*A man in a red hat is holding a shotgun in the air.*	*A dolphin fin is up in the water.*

nocaps.org

nocaps

- For societal impact and research challenge, why not aim for image captioning in the wild?
- Such a challenge should require transfer learning from multiple data sources
- nocaps can measure our progress
- Better object detectors and language models will help, but other challenges remain and become more explicit: e.g. visual saliency, entry-level categories, synonyms, grounding, etc.

nocaps.org

nocaps

- For societal impact and research challenge, why not aim for image captioning in the wild?
- Such a challenge should require transfer learning from multiple data sources
- nocaps can measure our progress
- Better object detectors and language models will help, but other challenges remain and become more explicit: e.g. visual saliency, entry-level categories, synonyms, grounding, etc.
- Full details in the arXiv paper, available via nocaps.org.

nocaps.org

Outline

Generating Visually-Grounded Language
(Image Captioning – Novel Object Captioning)

Understanding Visually-Grounded Language
(Vision-and-Language Navigation)

Future Work

From Static Images to Environments

From Static Images to Environments

- Tipping point for 3D reconstruction technology

From Static Images to Environments

- Tipping point for 3D reconstruction technology
- Untrained users producing millions of high quality 3D reconstructions

From Static Images to Environments

- Tipping point for 3D reconstruction technology
- Untrained users producing millions of high quality 3D reconstructions

From Static Images to Environments

- Tipping point for 3D reconstruction technology
- Untrained users producing millions of high quality 3D reconstructions

Image source: Matterport

From Static Images to Environments

Slide Credit: Dhruv Batra

From Static Images to Environments

Tasks, Metrics
& Algorithms

Environments

Raw Data

Slide Credit: Dhruv Batra

From Static Images to Environments

Tasks, Metrics & Algorithms

EmbodiedQA

Language grounding (Chaplot et al., 2017, Hermann & Hill et al., 2017)

Interactive QA (Gordon et al., 2018)

Vision-Language Navigation (Anderson et al., 2018)

Visual Navigation (Zhu & Gordon et al., 2017, Savva et al., 2017, Wu et al., 2017)

Environments

House3D (Wu et al., 2017)

AI2-THOR (Kolve et al., 2017)

MINOS (Savva et al., 2017)

Gibson (Zamir et al., 2018)

CHALET (Yan et al., 2018)

HoME (Brodeur et al., 2018)

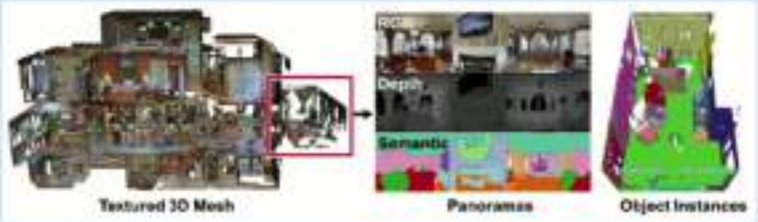
VirtualHome (Puig et al., 2018)

AdobeIndoorNav (Mo et al., 2018)

Matterport3DSim (Anderson et al., 2018)

Raw Data

SUNCG (Song et al., 2017)



Matterport3D (Chang et al., 2017)

Stanford 2D-3D-S (Armeni et al., 2017)

Slide Credit: Dhruv Batra

From Static Images to Environments

>= 2017 (!)

Tasks, Metrics & Algorithms

EmbodiedQA

Language grounding (Chaplot et al., 2017, Hermann & Hill et al., 2017)

Interactive QA (Gordon et al., 2018)

Vision-Language Navigation (Anderson et al., 2018)

Visual Navigation (Zhu & Gordon et al., 2017, Savva et al., 2017, Wu et al., 2017)

Environments

House3D (Wu et al., 2017)

AI2-THOR (Kolve et al., 2017)

MINOS (Savva et al., 2017)

Gibson (Zamir et al., 2018)

CHALET (Yan et al., 2018)

HoME (Brodeur et al., 2018)

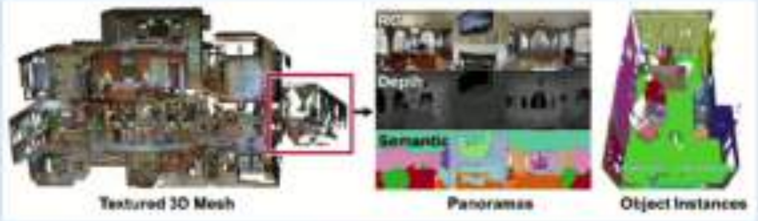
VirtualHome (Puig et al., 2018)

AdobeIndoorNav (Mo et al., 2018)

Matterport3DSim (Anderson et al., 2018)

Raw Data

SUNCG (Song et al., 2017)



Matterport3D (Chang et al., 2017)

Stanford 2D-3D-S (Armeni et al., 2017)

Slide Credit: Dhruv Batra

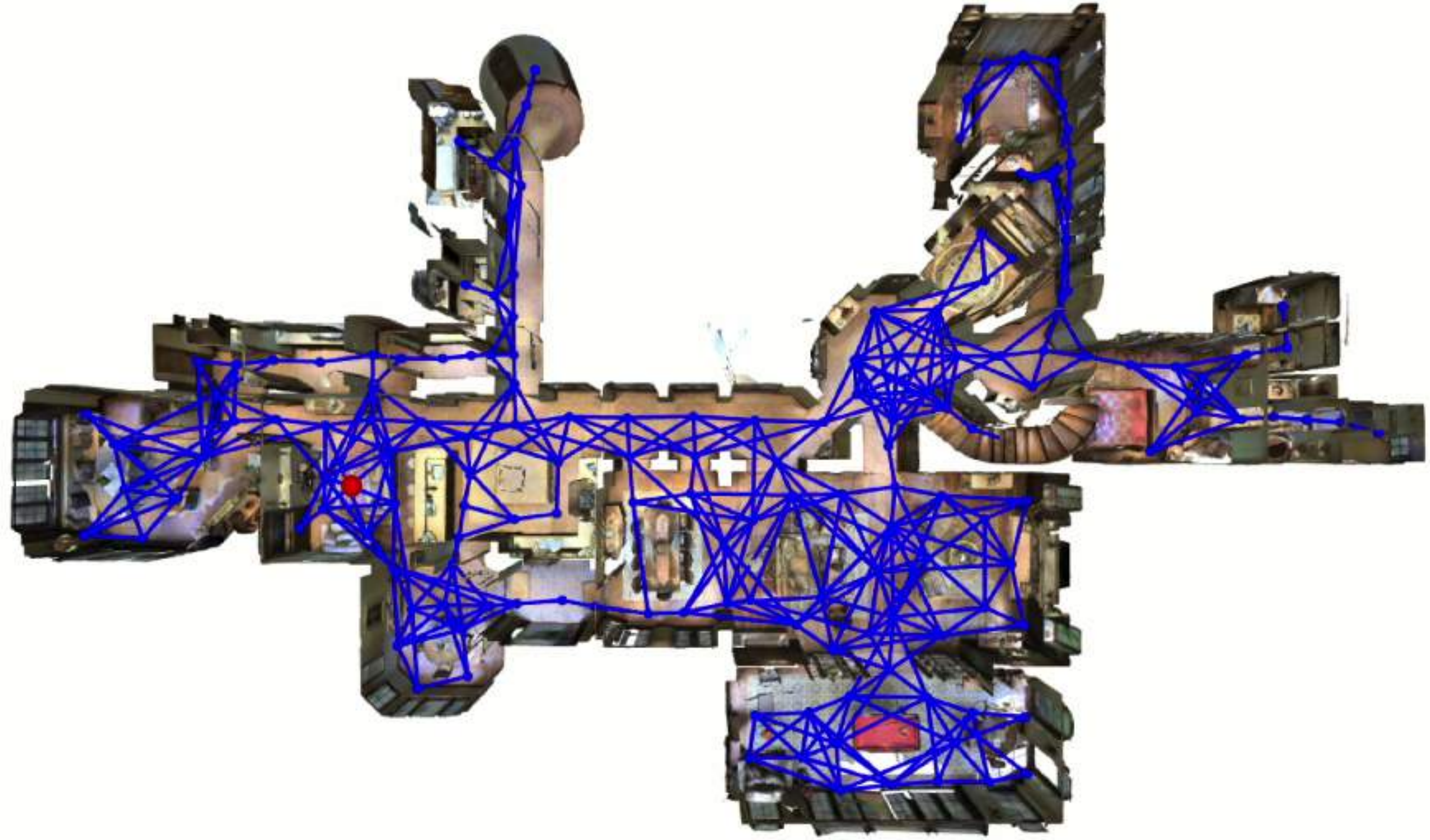
Matterport3D Simulator

Simulator for embodied visual agents, based on the Matterport3D dataset [*Chang et al.* 3DV 2017]:

- 10,800 panoramic RGB-D images
- Covering 90 buildings
- High visual diversity
- Additionally includes textured 3D meshes and object annotations

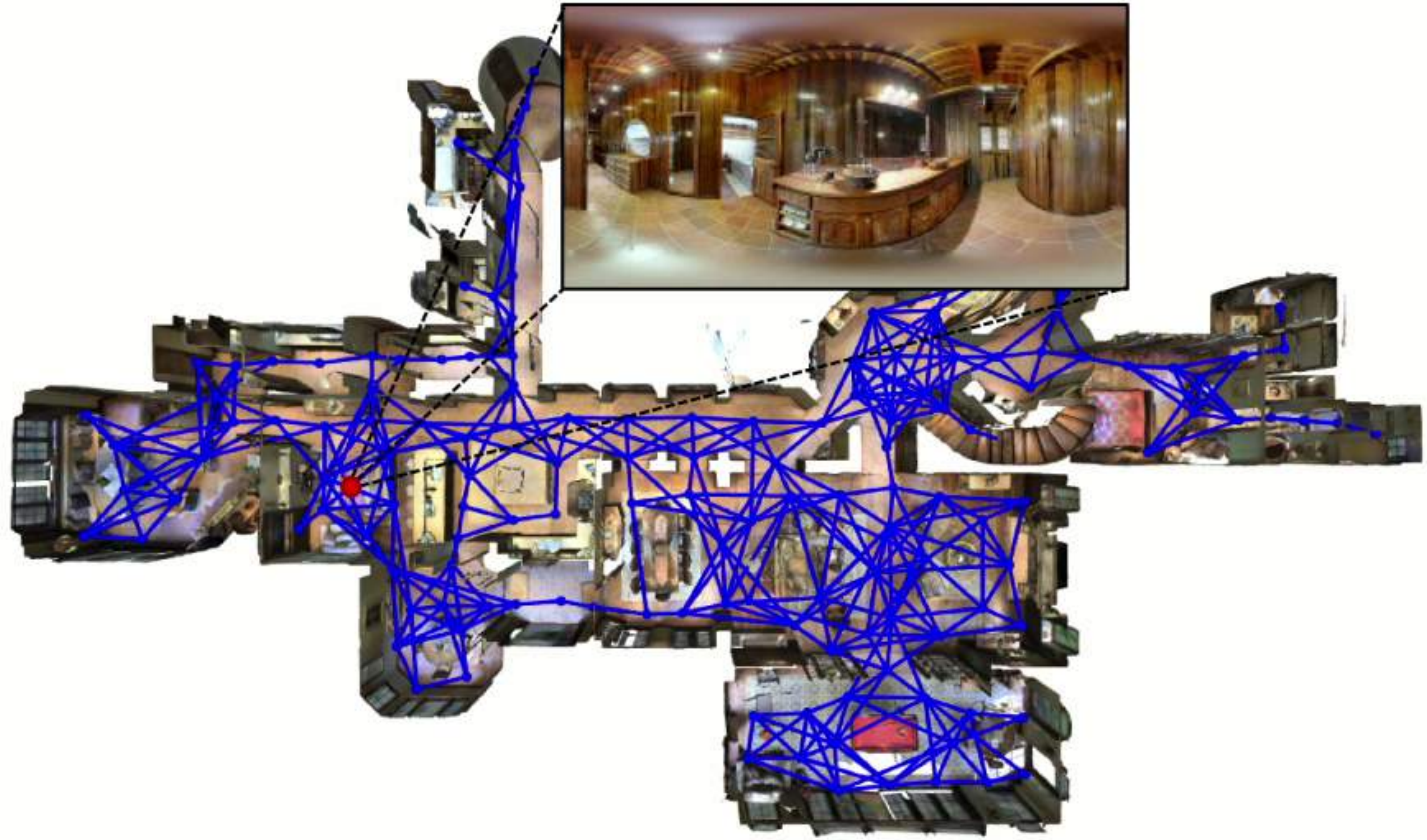
[*Chang et al.* 3DV 2017]

Matterport3D Simulator



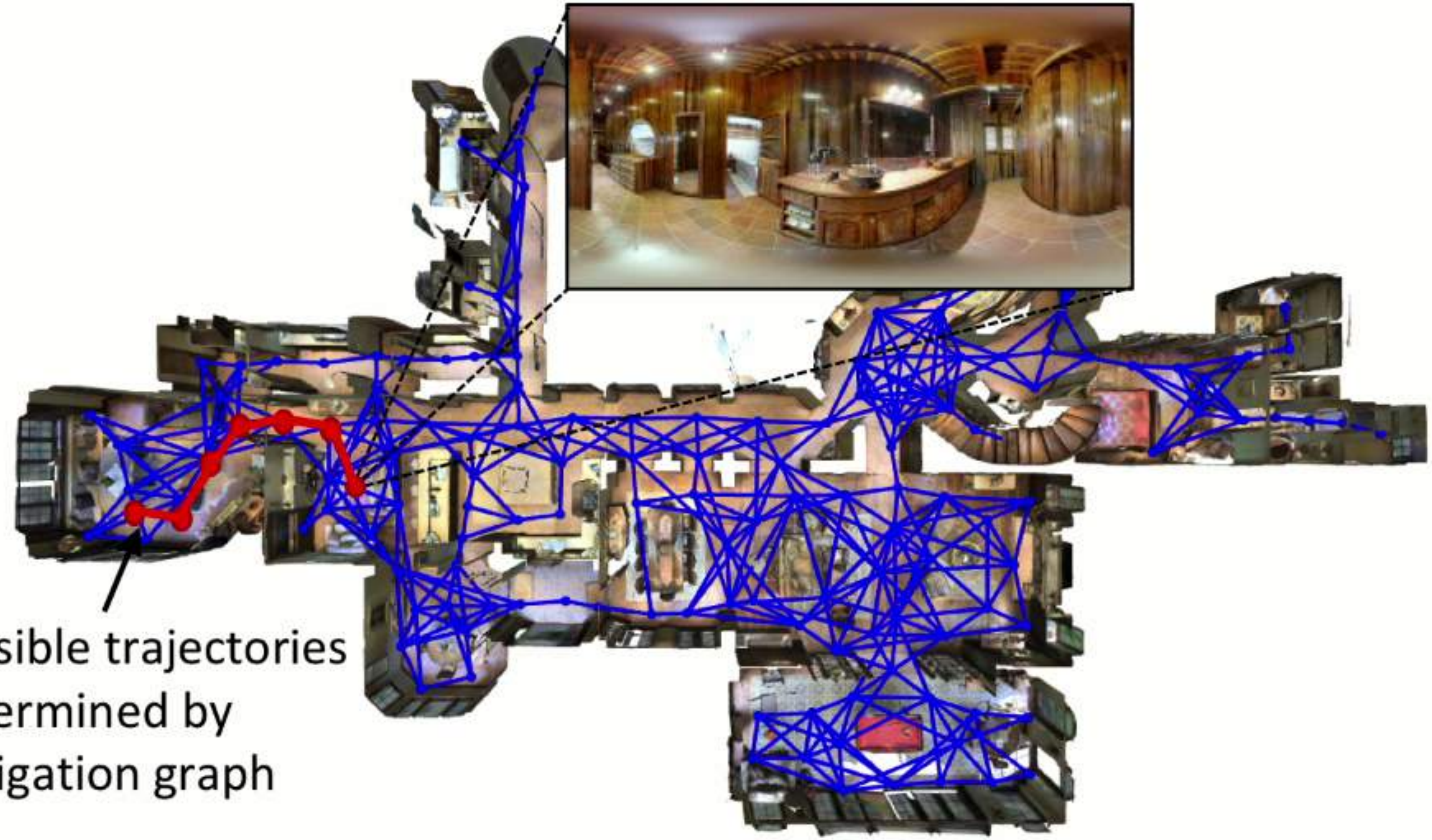
[*Anderson et al. CVPR* 2018]

Matterport3D Simulator



[*Anderson et al. CVPR* 2018]

Matterport3D Simulator



Feasible trajectories determined by navigation graph

[*Anderson et al. CVPR* 2018]

Room-to-Room (R2R) Dataset

- Vision-and-Language Navigation (VLN) task: given natural language instructions, find the goal location

Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and a table. Wait by the moose antlers hanging on the wall.

[*Anderson et al. CVPR* 2018]

Room-to-Room (R2R) Dataset

- Vision-and-Language Navigation (VLN) task: given natural language instructions, find the goal location
- Sampled 7,189 shortest paths between locations (mostly) in different rooms

Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and a table. Wait by the moose antlers hanging on the wall.

[*Anderson et al. CVPR* 2018]

Room-to-Room (R2R) Dataset

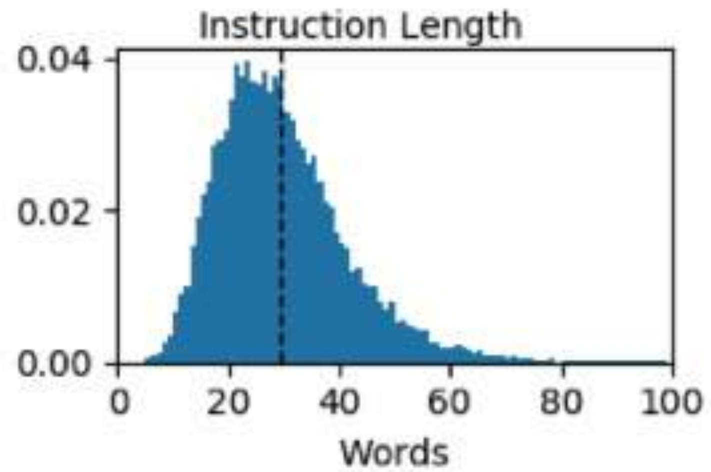
- Vision-and-Language Navigation (VLN) task: given natural language instructions, find the goal location
- Sampled 7,189 shortest paths between locations (mostly) in different rooms
- Collected 21,567 instructions using AMT and a WebGL simulator interface

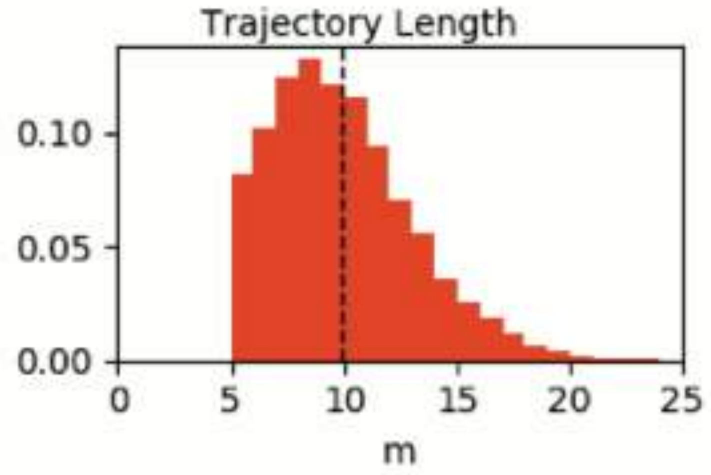
Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and a table. Wait by the moose antlers hanging on the wall.

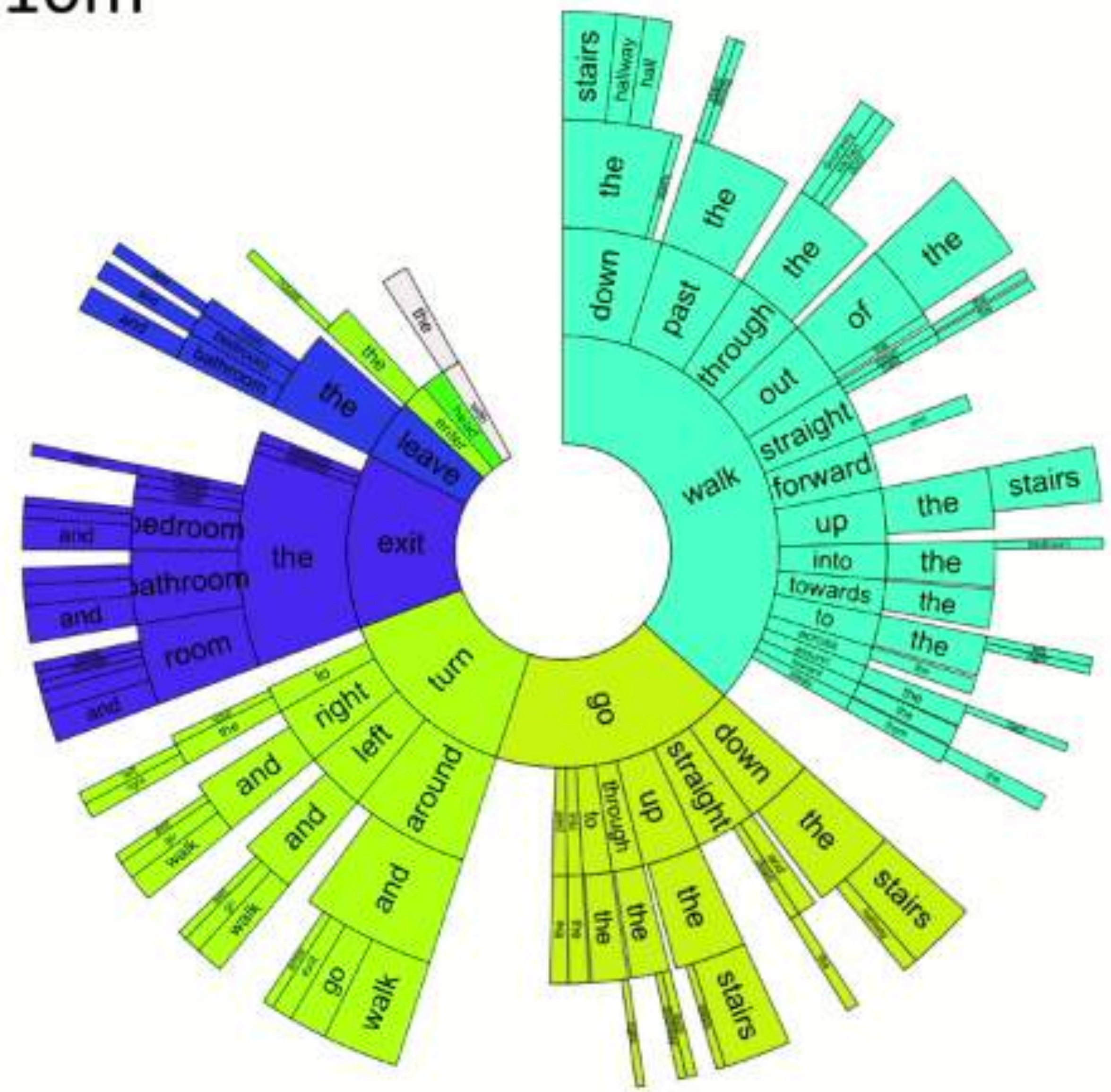
[*Anderson et al. CVPR* 2018]

Room-to-Room (R2R) Dataset

- Average instruction length is 29 words
- Average trajectory length is 10m







[*Anderson et al. CVPR* 2018]

Examples

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

[*Anderson et al. CVPR* 2018]

SOTA on R2R Dataset

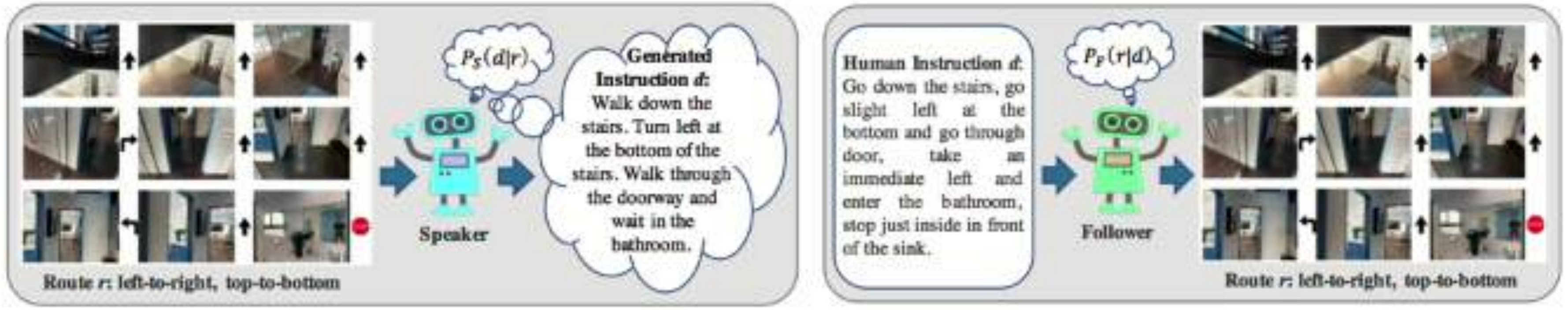
- Test server on EvalAI
- 28 submissions since Sept 2018

Rank	Participant team	length	error	oracle success	success	spl	Last submission at
1	human	11.85	1.61	0.90	0.86	0.76	7 months ago
3	Back Translation with Environmental Dropout (exploring unseen environments before testing)	9.79	3.97	0.70	0.64	0.61	4 months ago
8	Reinforced Cross-Modal Matching + SIL (exploring unseen environments before testing)	9.48	4.21	0.67	0.60	0.59	5 months ago
13	Back Translation with Environmental Dropout (no beam search)	11.66	5.23	0.59	0.51	0.47	4 months ago
16	ALTR	10.27	5.49	0.56	0.48	0.45	1 month ago
18	naive	10.42	5.64	0.53	0.47	0.43	1 month ago
10	Tactical Rewind - short	22.08	5.14	0.64	0.54	0.41	5 months ago

What's driving progress?

- Pragmatic Reasoning
- Stopping Models
- Reasoning about Backtracking
- Data Aug. / Back-Translation
- Environmental Dropout
- Beam Search for robots?!
- Lacking algorithmic and geometric priors

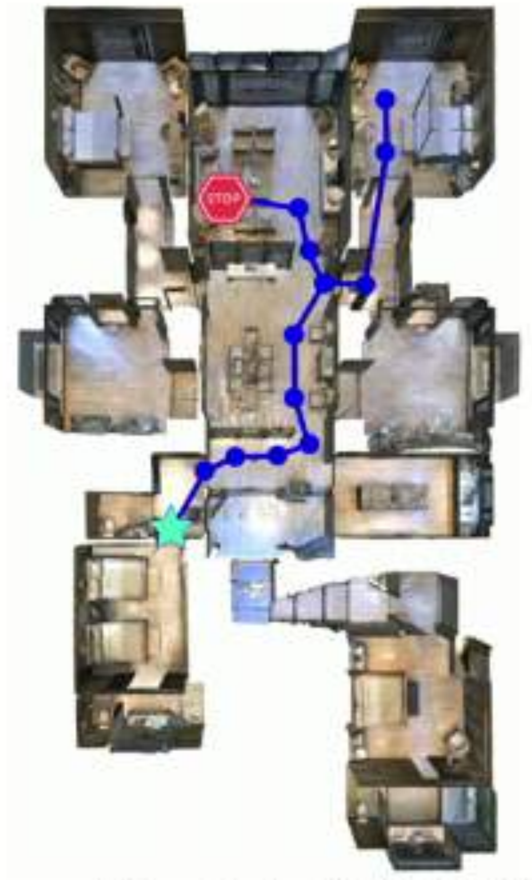
What's driving progress?



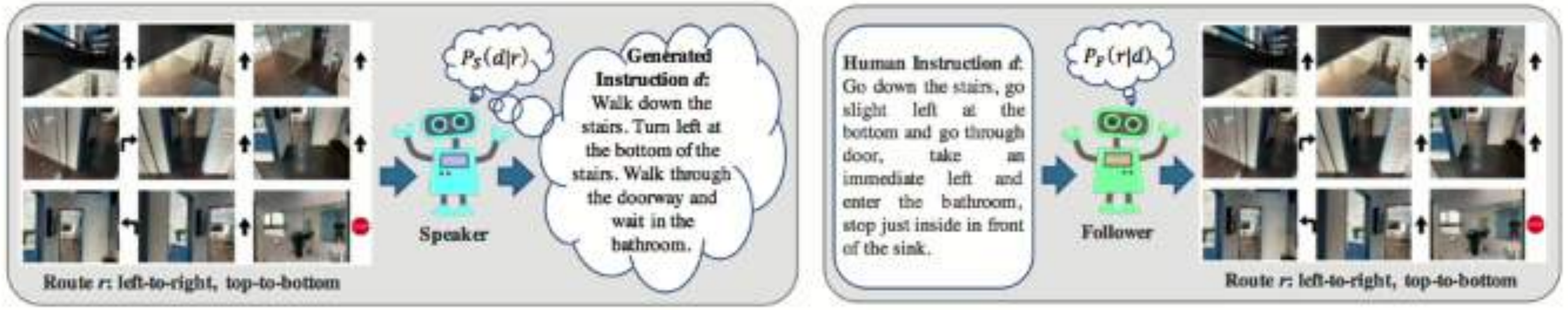
[*Fried et al. NeurIPS* 2018]

- Pragmatic Reasoning
- Stopping Models
- Reasoning about Backtracking
- Data Aug. / Back-Translation
- Environmental Dropout
- Beam Search for robots?!
- Lacking algorithmic and geometric priors

What's driving progress?



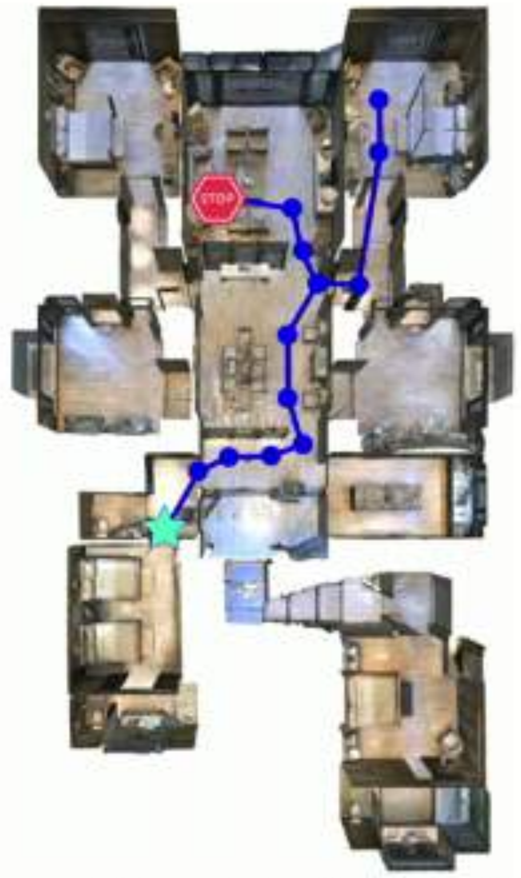
[*Ke et al. CVPR* 2019]



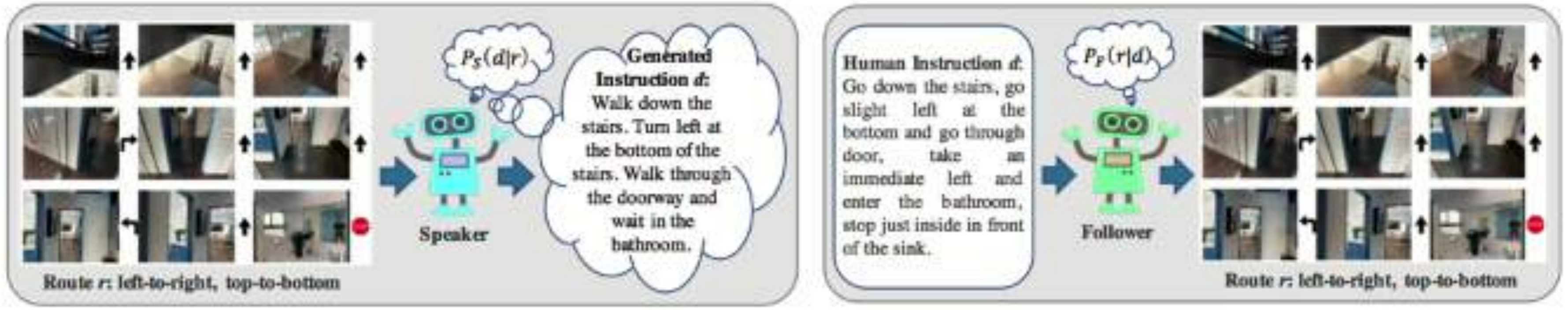
[*Fried et al. NeurIPS* 2018]

- Pragmatic Reasoning
- Stopping Models
- Reasoning about Backtracking
- Data Aug. / Back-Translation
- Environmental Dropout
- Beam Search for robots?!
- Lacking algorithmic and geometric priors

What's driving progress?



[*Ke et al. CVPR* 2019]



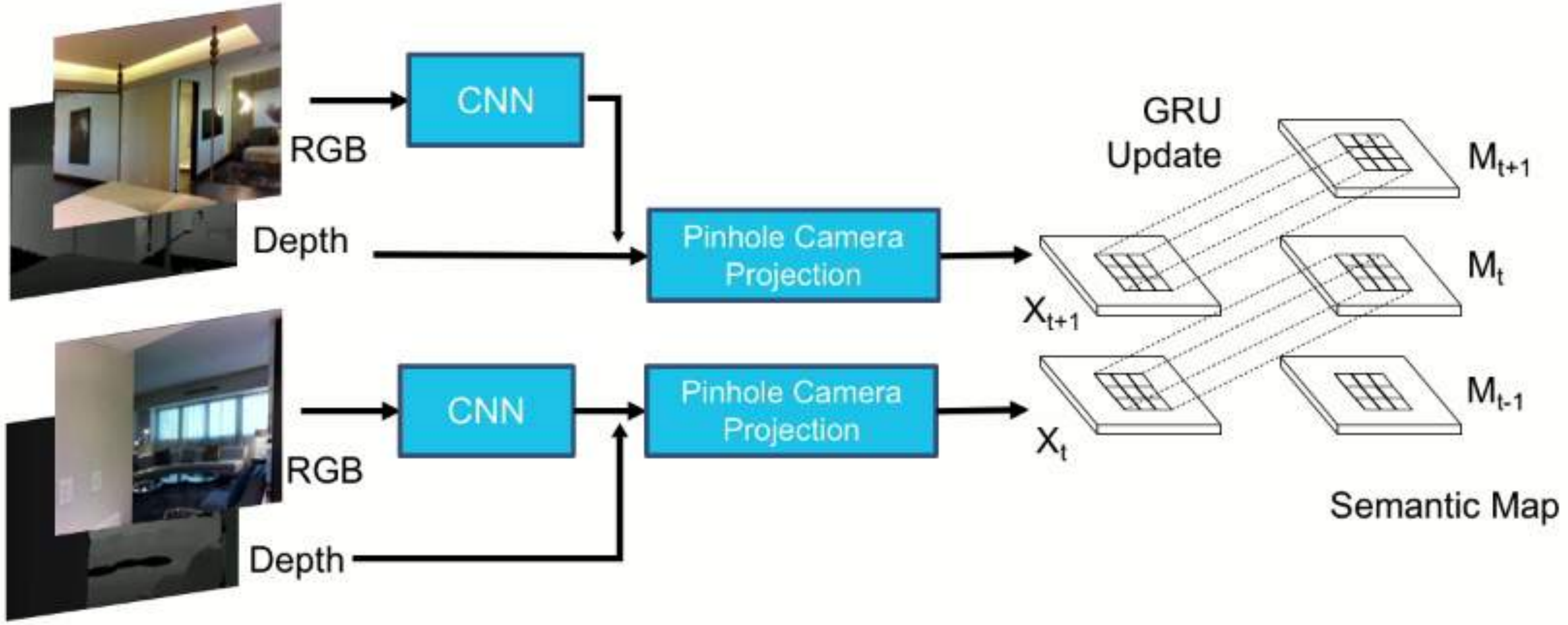
[*Fried et al. NeurIPS* 2018]

[*Tan et al. NAACL* 2019]

- Pragmatic Reasoning
- Stopping Models
- Reasoning about Backtracking
- Data Aug. / Back-Translation
- Environmental Dropout
- Beam Search for robots?!
- Lacking algorithmic and geometric priors

Incorporating Geometry

- Extended Matterport3D Simulator to provide depth
- Construct a semantic metric map by projecting CNN features to the ground plane with a pinhole camera model



Semantic Spatial Maps – refer [*Gordon et al.* CVPR 2018, *Henriques et al.* CVPR 2018, etc]

How to Use a Map?

Instruction: Turn and walk out of the bathroom into the hallway. Walk through the door into the room with shelves and a sink. Continue through the room into the kitchen area and walk past the stove and sink.

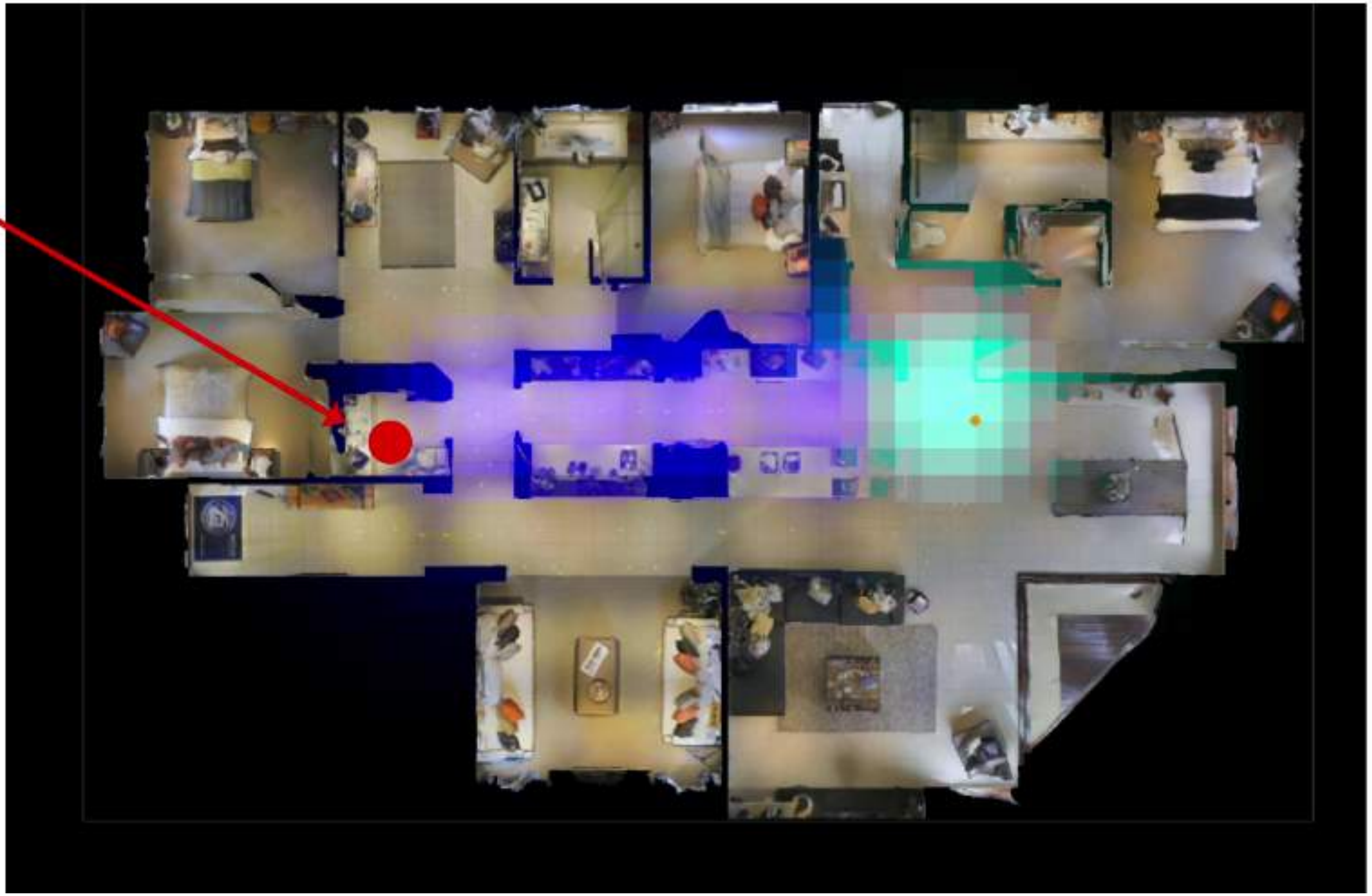
How to Use a Map?

Instruction: Turn and walk out of the **bathroom** into the **hallway**. Walk through the door into the room with **shelves and a sink**. Continue through the room into the **kitchen area** and walk past the **stove and sink**.

How to Use a Map?

Instruction: Turn and walk out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

Start



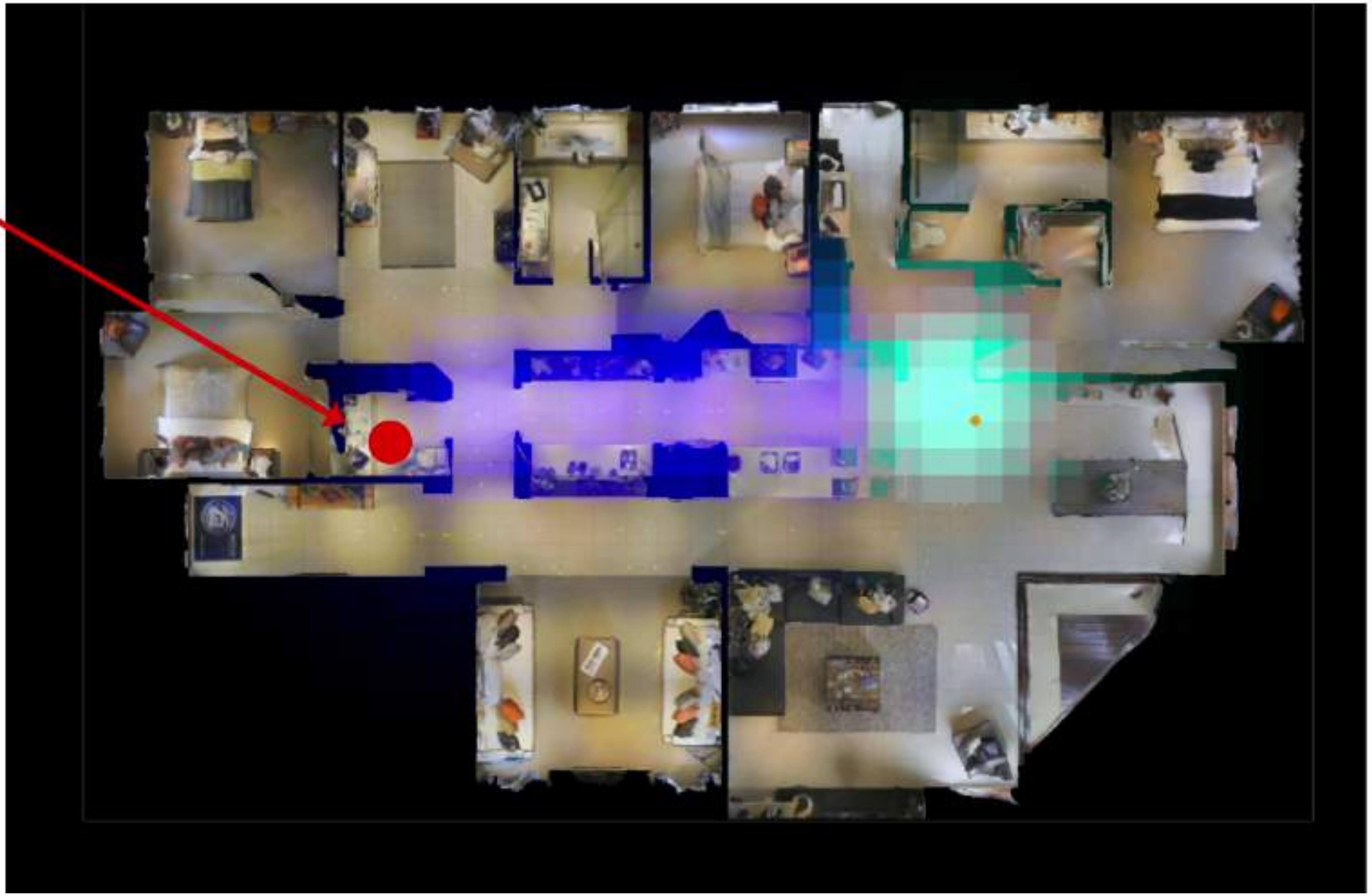
How to Use a Map?

ACTIONS

OBSERVATIONS

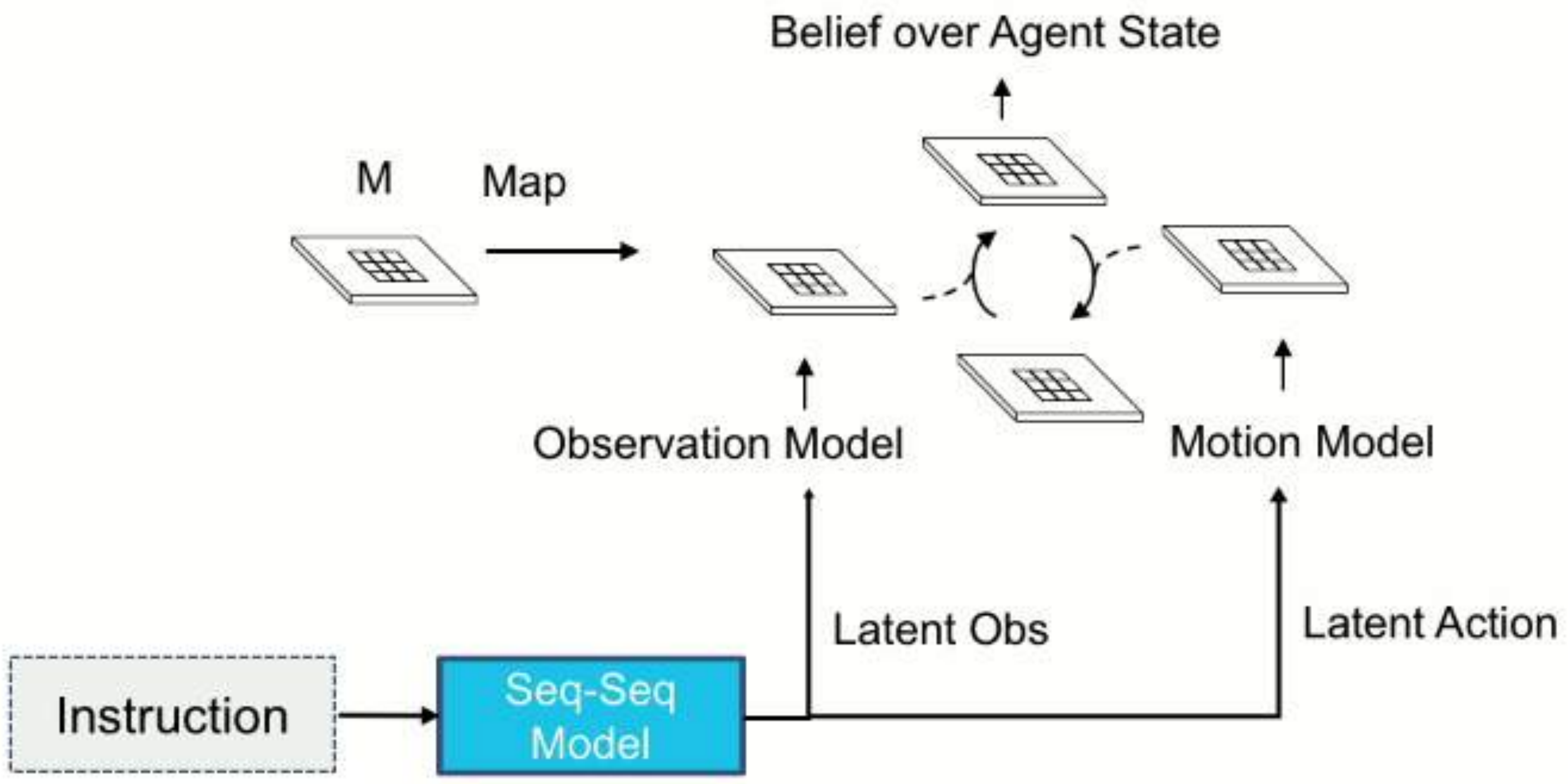
Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

Start



Bayes Filter

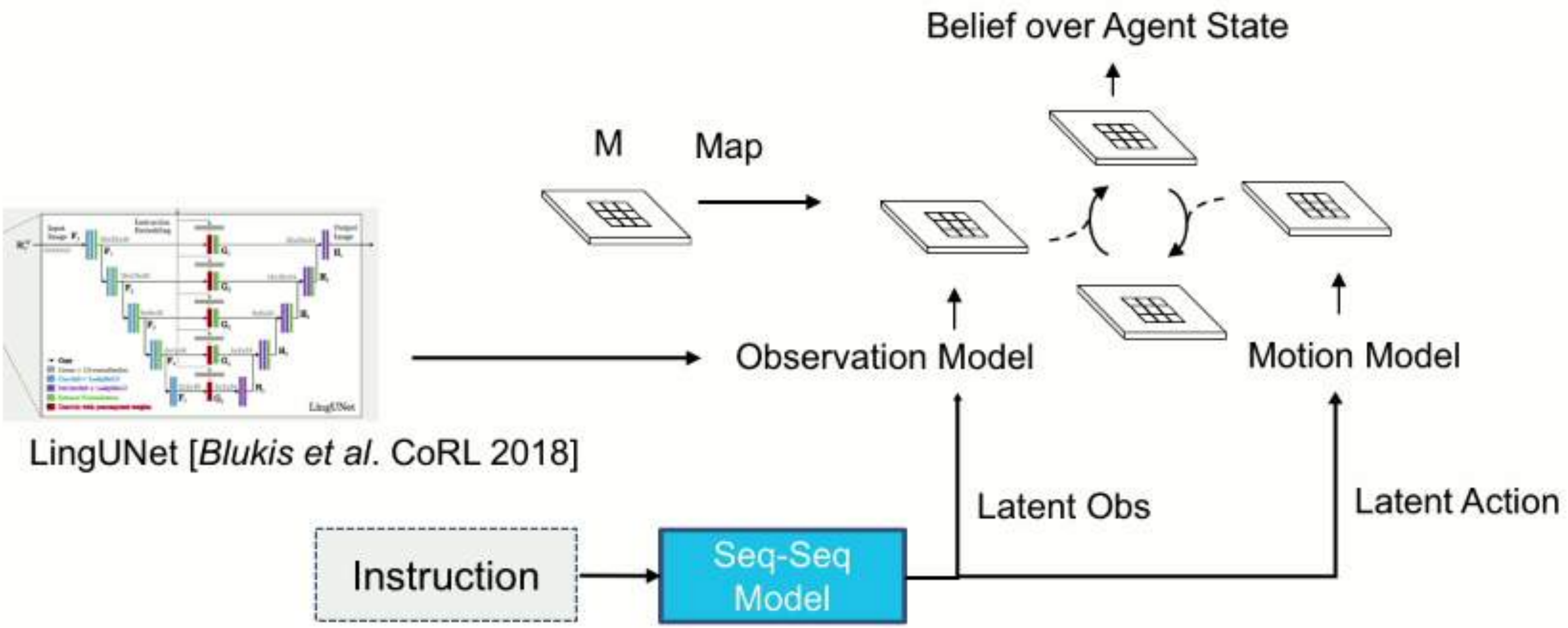
- A visually-grounded navigation instruction can be interpreted as a sequence of observations and actions.
- We have a map. Let’s track the human demonstrator!



Differentiable Bayesian Filters – refer [*Jonschkowski et al.* RSS 2018, *Karkus et al.* CoRL 2018]

Bayes Filter

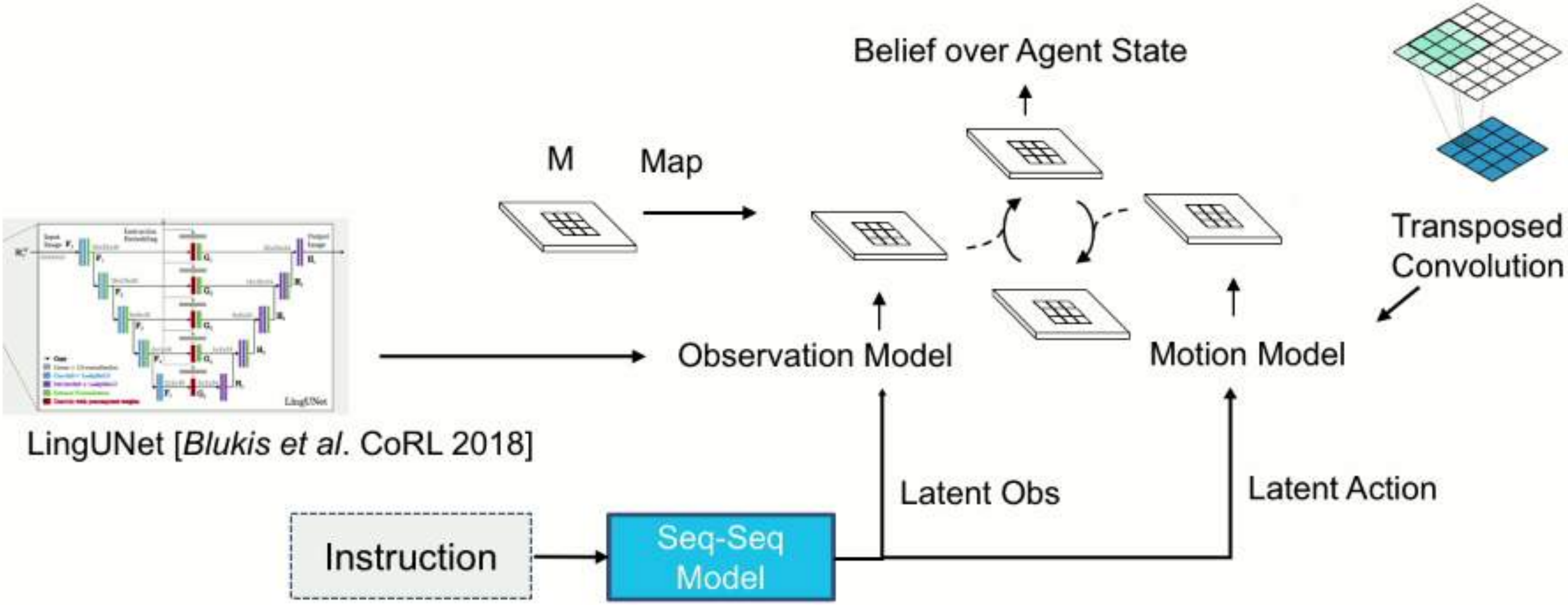
- A visually-grounded navigation instruction can be interpreted as a sequence of observations and actions.
- We have a map. Let's track the human demonstrator!



Differentiable Bayesian Filters – refer [*Jonschkowski et al*. RSS 2018, *Karkus et al*. CoRL 2018]

Bayes Filter

- A visually-grounded navigation instruction can be interpreted as a sequence of observations and actions.
- We have a map. Let's track the human demonstrator!



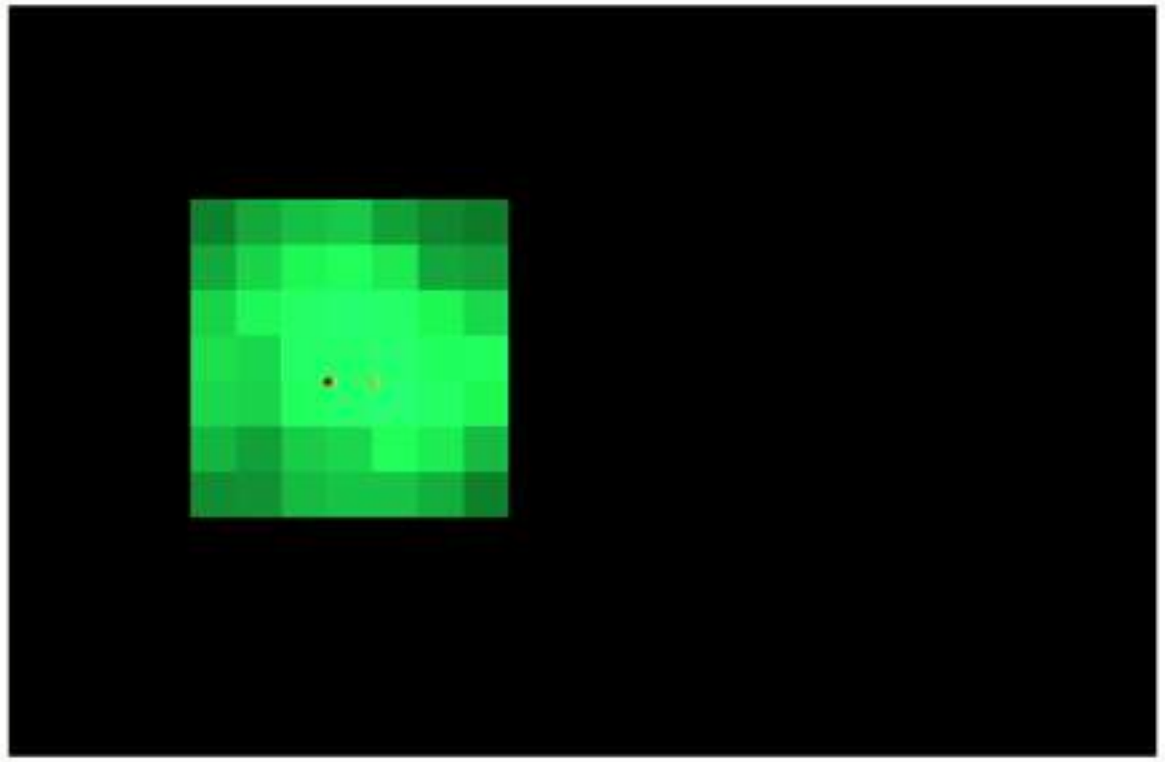
Differentiable Bayesian Filters – refer [*Jonschkowski et al.* RSS 2018, *Karkus et al.* CoRL 2018]

Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

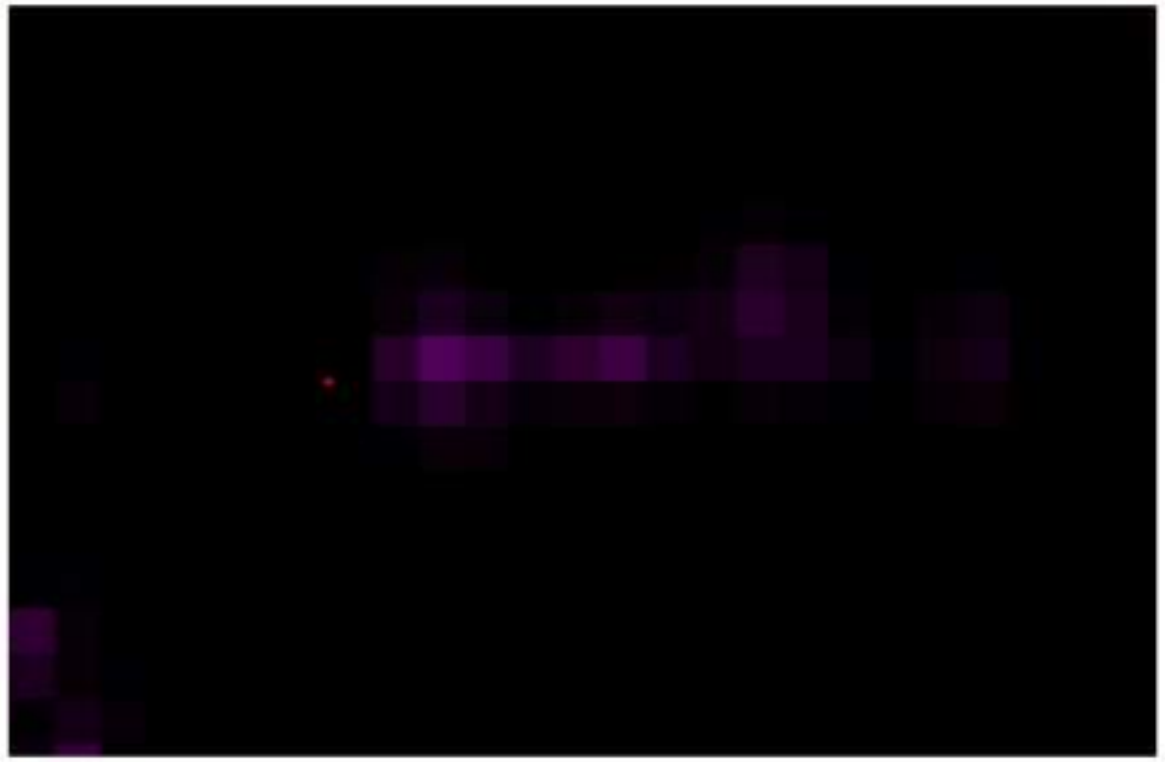


ACTIONS

Example

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

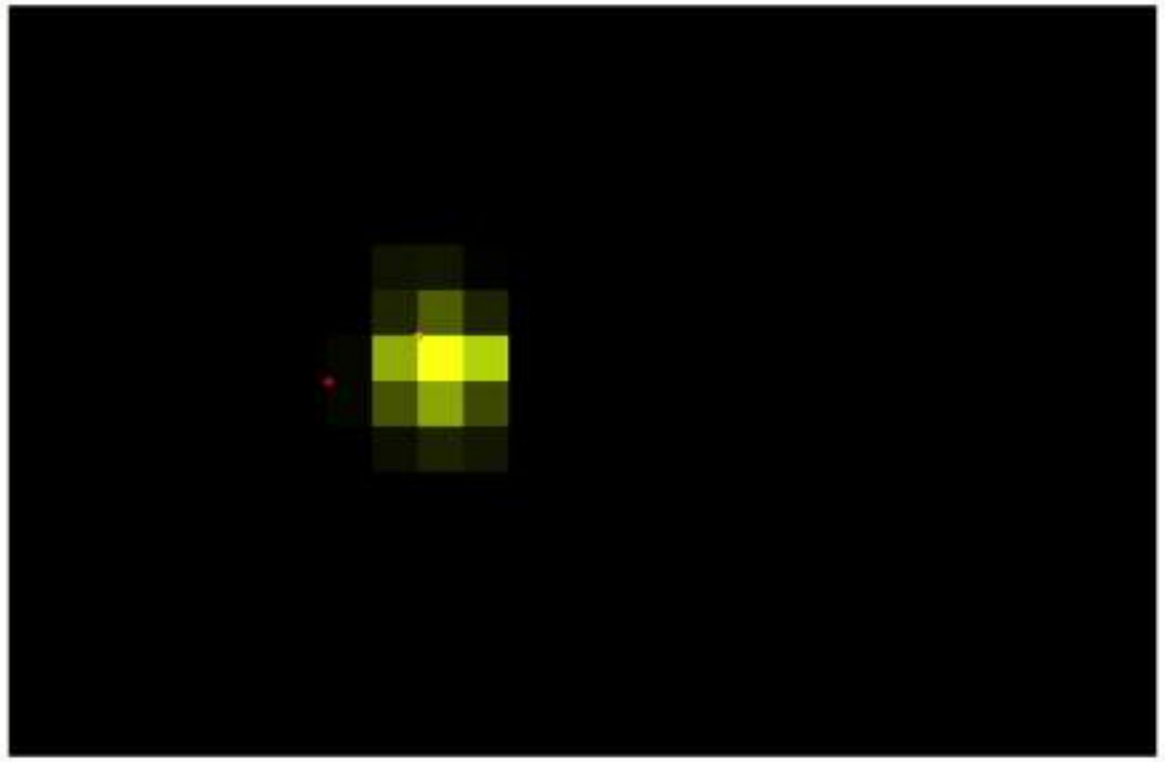


Example

ACTIONS

OBSERVATIONS

Instruction: Turn and walk out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

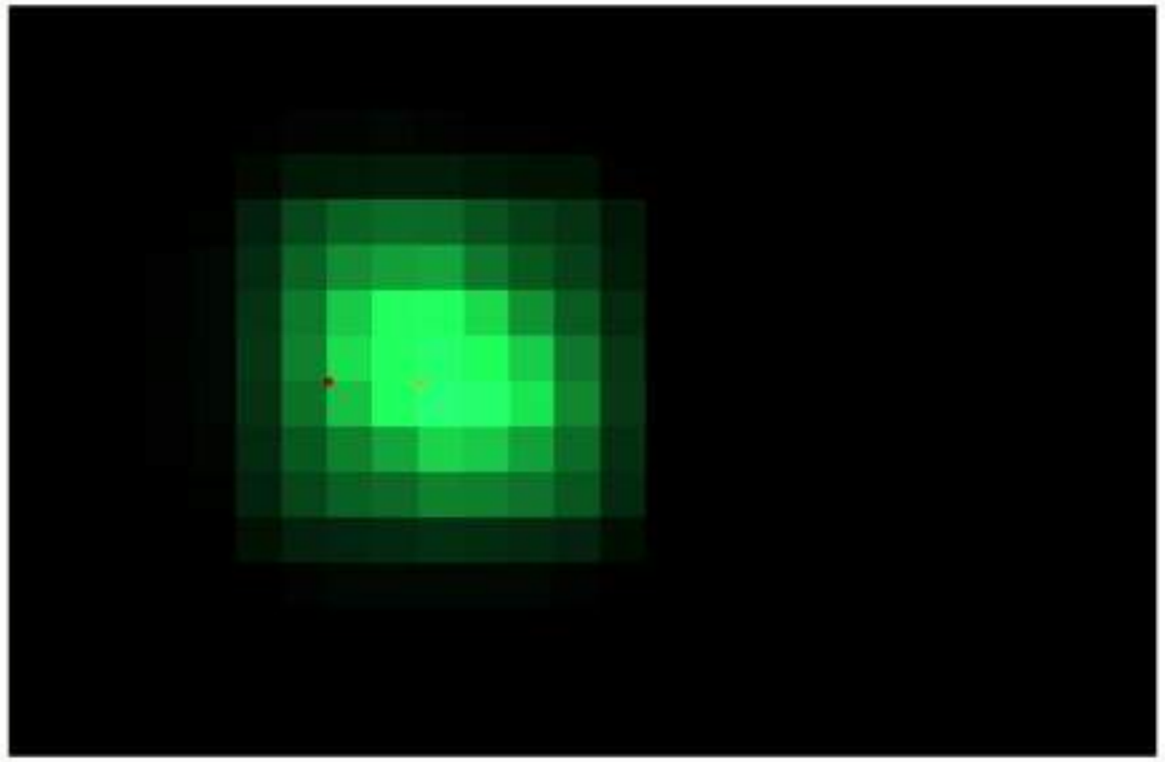


Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

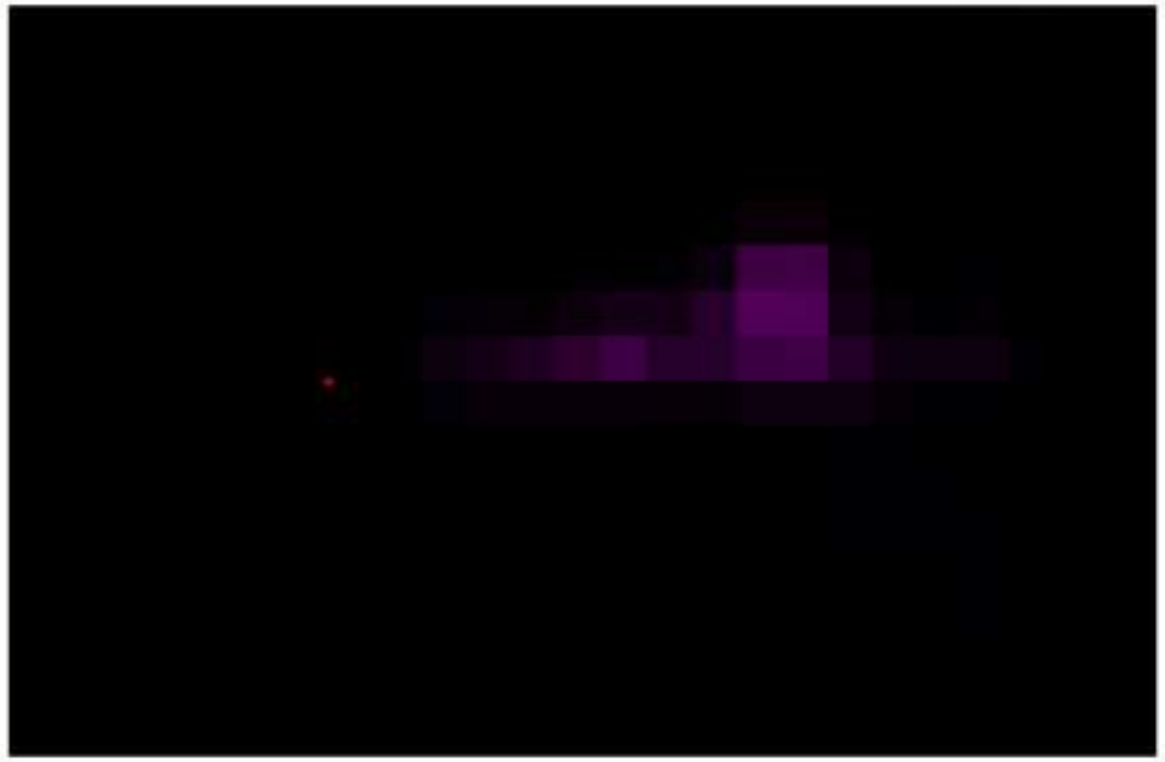


Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

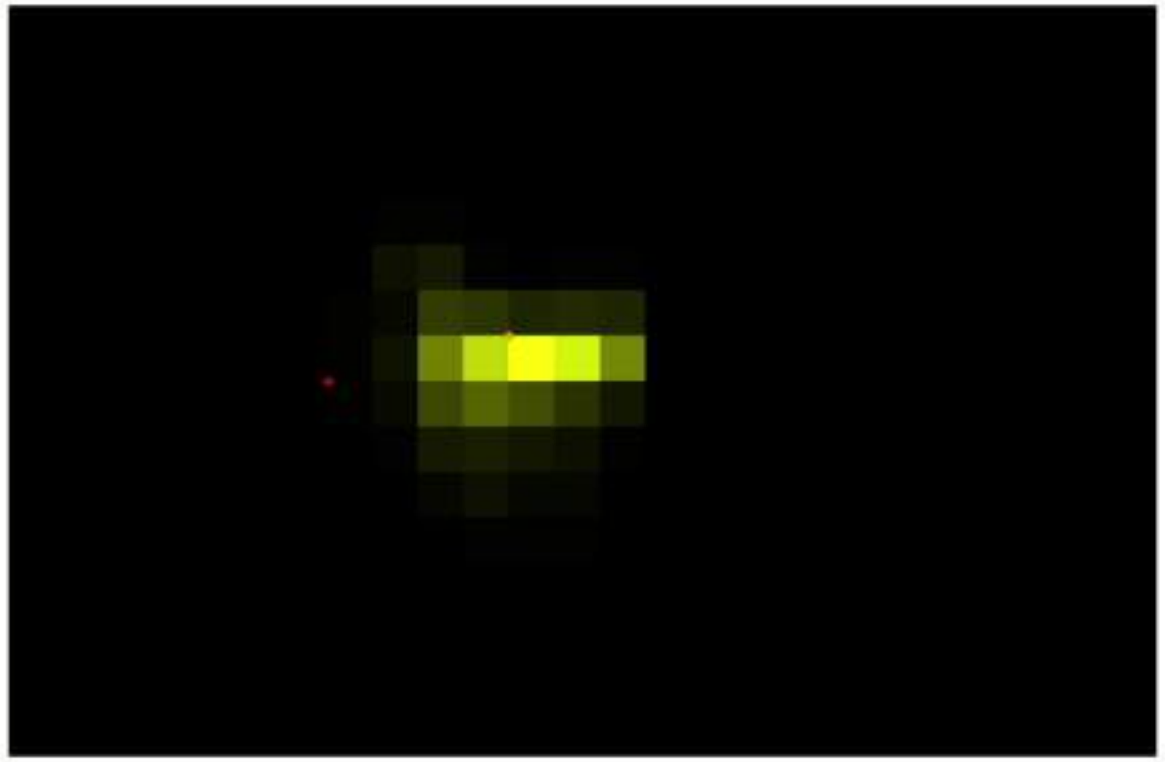


Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

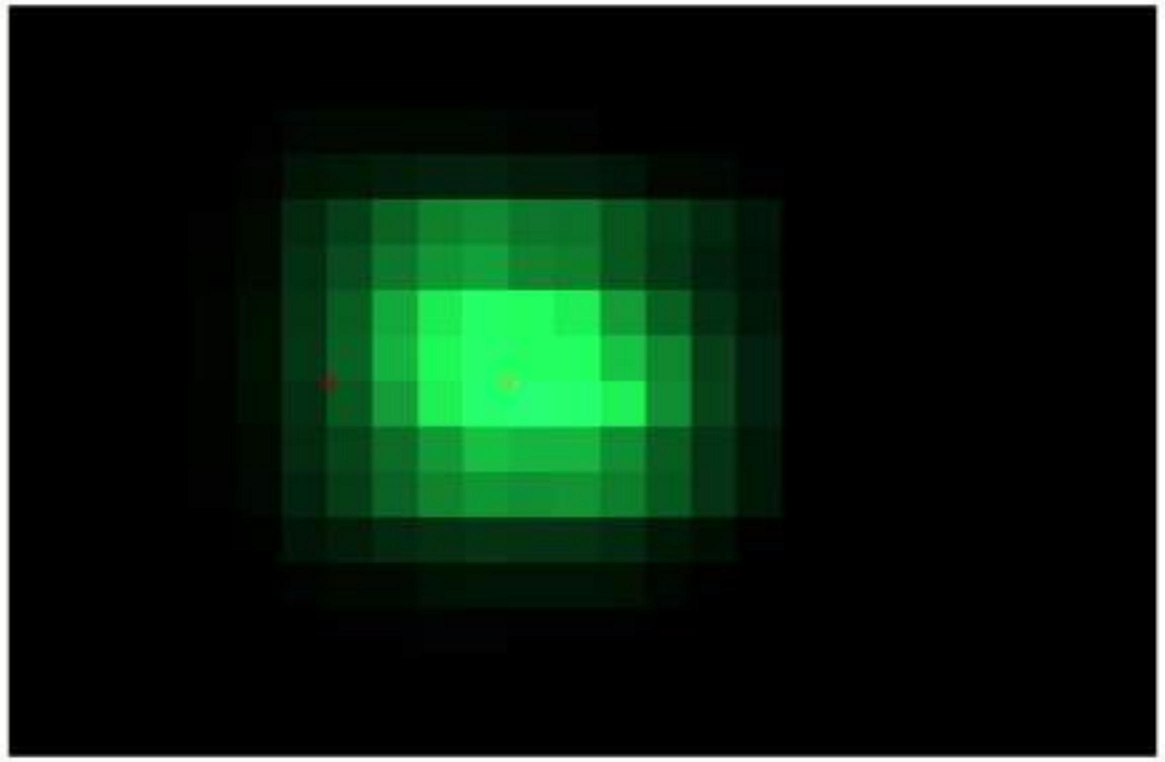


Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

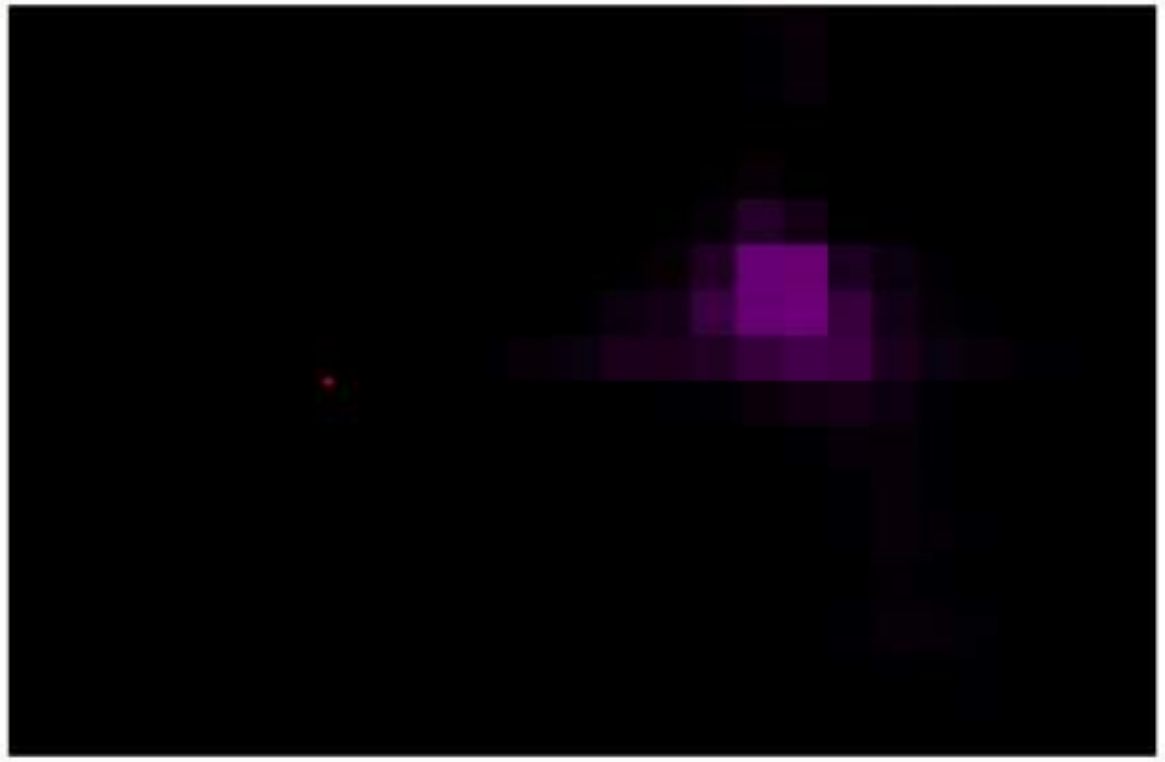


Example

ACTIONS

OBSERVATIONS

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

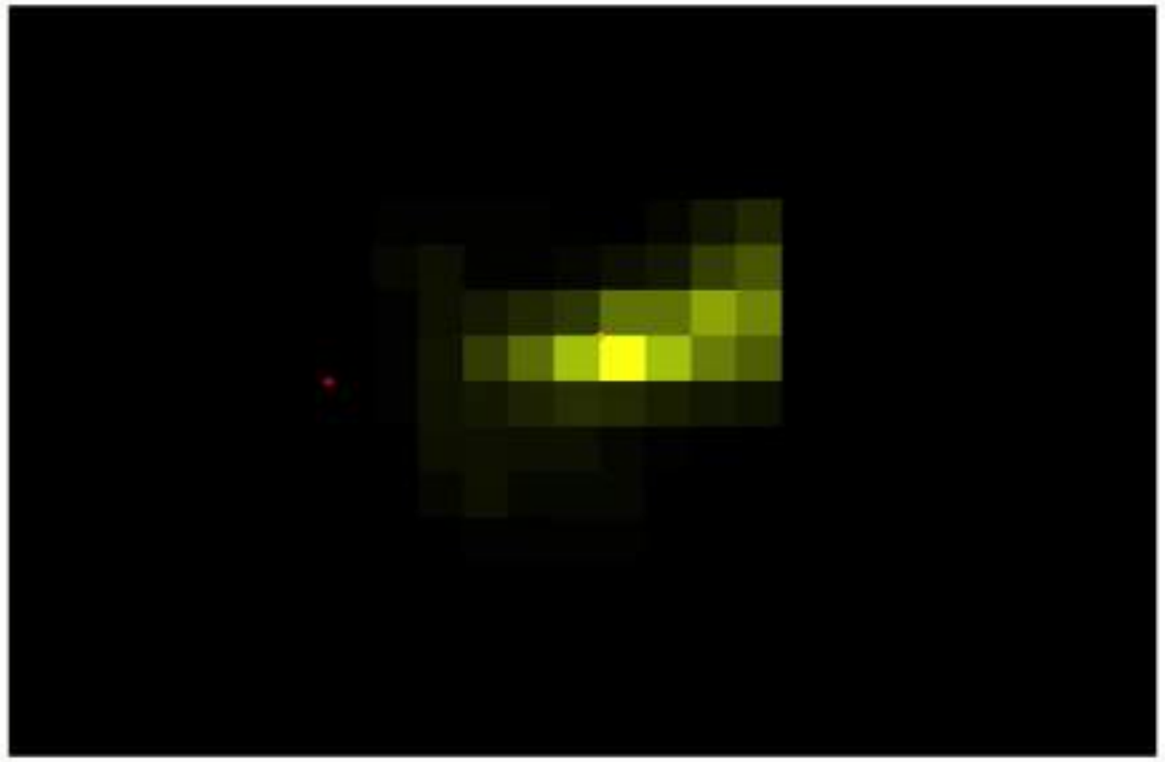


Example

ACTIONS

OBSERVATIONS

Instruction: Turn and walk out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.

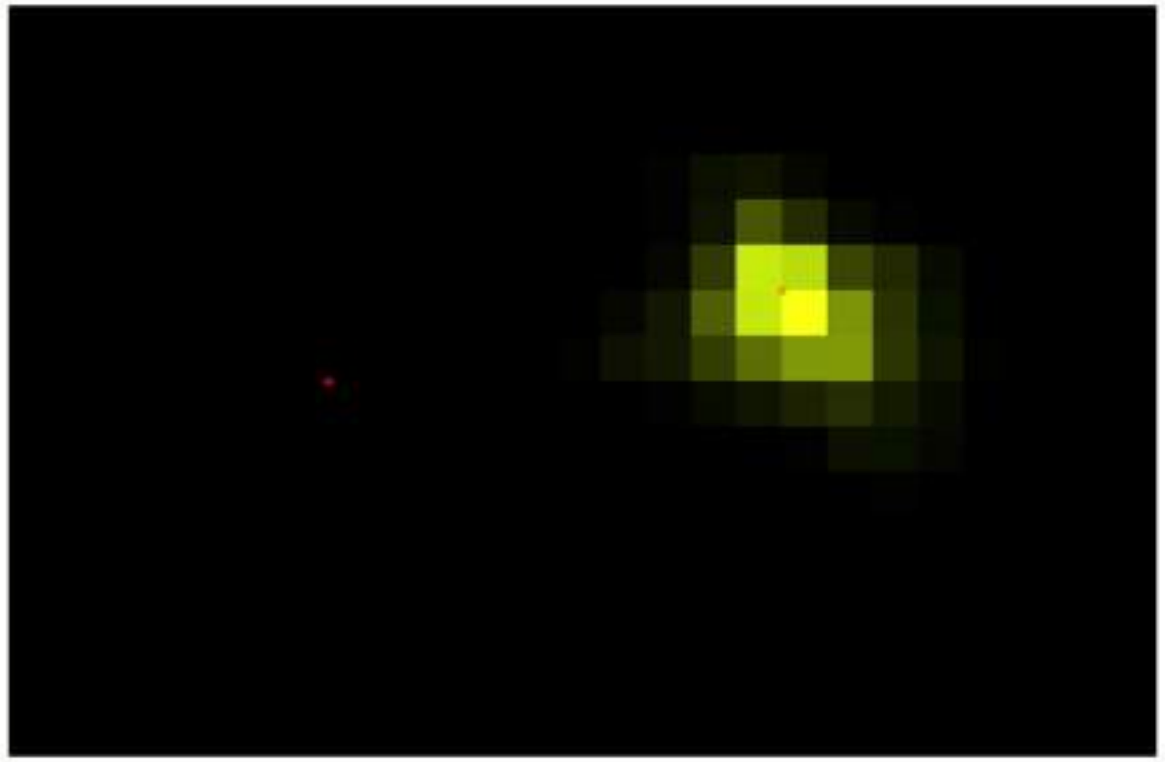


Example

ACTIONS

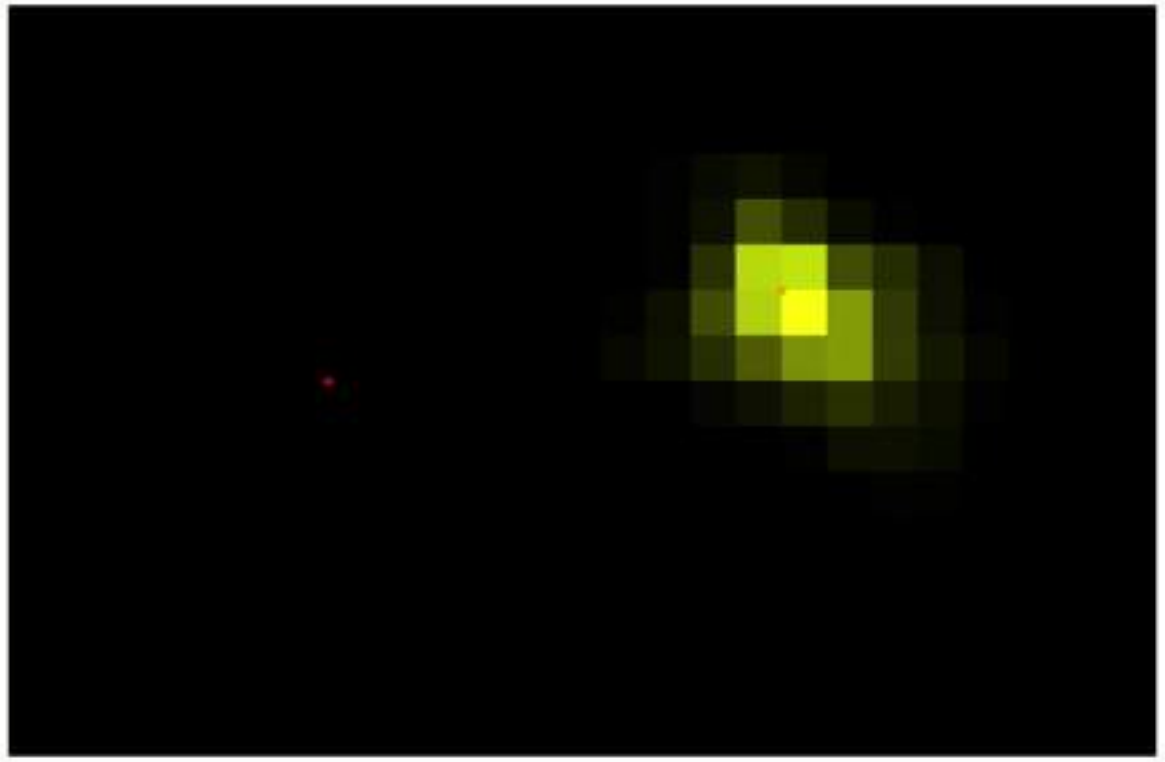
OBSERVATIONS

Instruction: Turn and walk out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.



Example

Instruction: **Turn and walk** out of the **bathroom** into the **hallway**. **Walk through** the door into the room with **shelves and a sink**. **Continue through** the room into the **kitchen area** and **walk past** the **stove and sink**.



- Future extensions: heading state, add policy for full VLN model

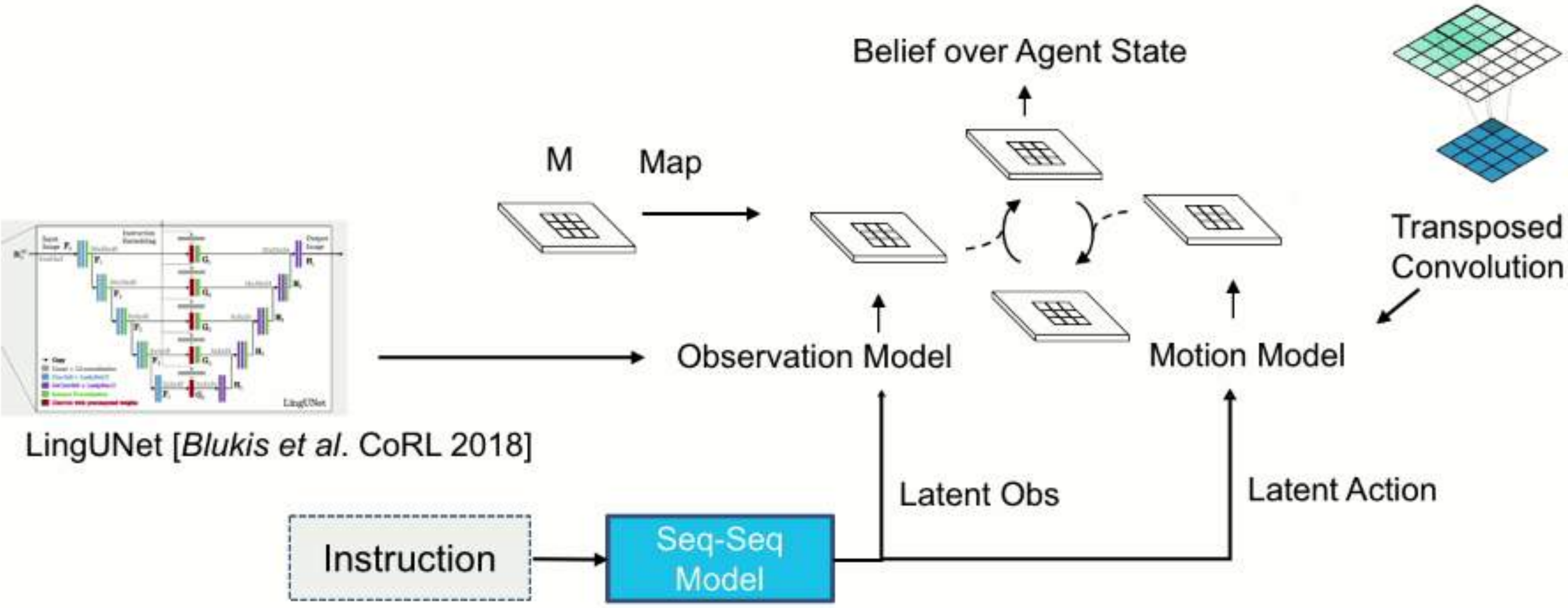
Experiments

- Fixed trajectories, 70% towards goal, 30% random
- Predict the goal location for val-unseen

Timestep	**0**	**1**	**2**	**3**	**4**	**5**	**Avg**
Map Area (m^2)	32.9	47.0	57.9	66.8	74.9	82.2	60.3
Goal Seen (%)	10.98	18.85	30.21	45.83	57.66	67.91	38.57
Success Rate (<3m):							
Handcoded	18.8	21.2	23.2	24.3	25.7	26.5	23.3
LingUNet only	23.8	25.4	34.4	41.9	50.3	56.6	38.7
Filter	**31.7**	**37.6**	**44.2**	**49.4**	**56.2**	**60.0**	**46.5**

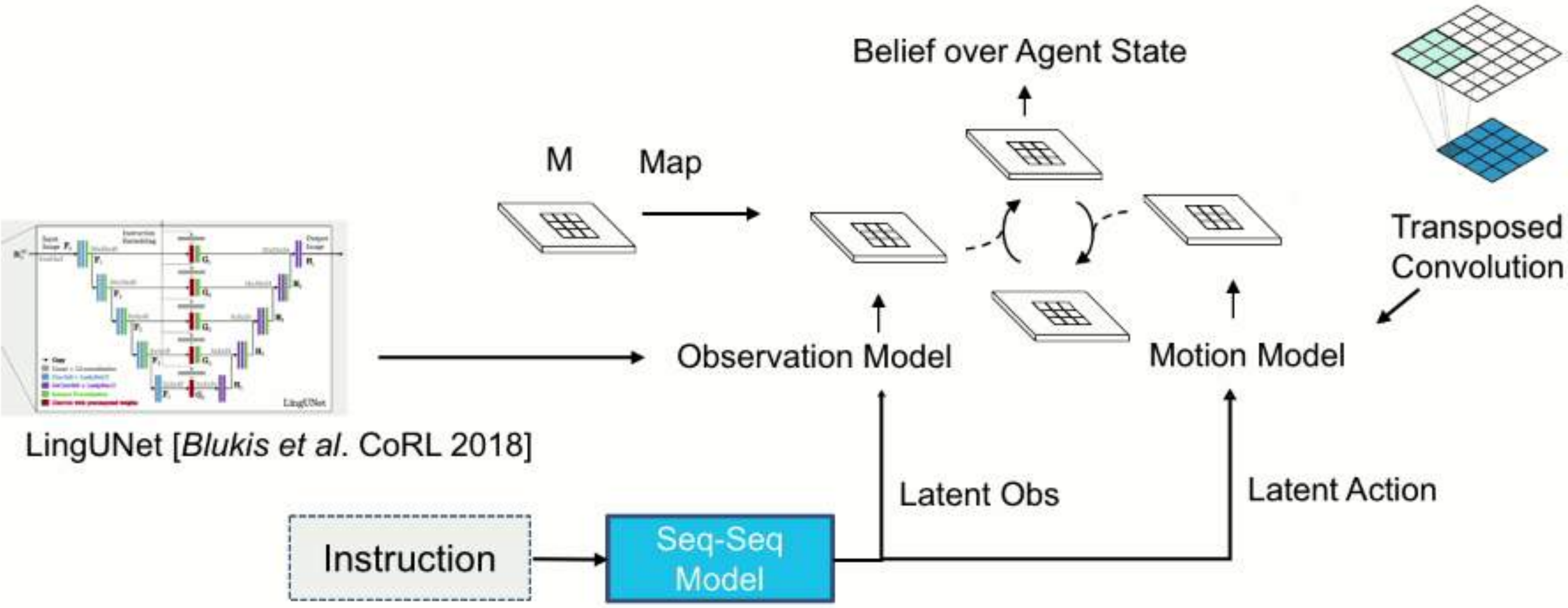
VLN + Bayes Filter

- Incorporates geometric priors via camera projection



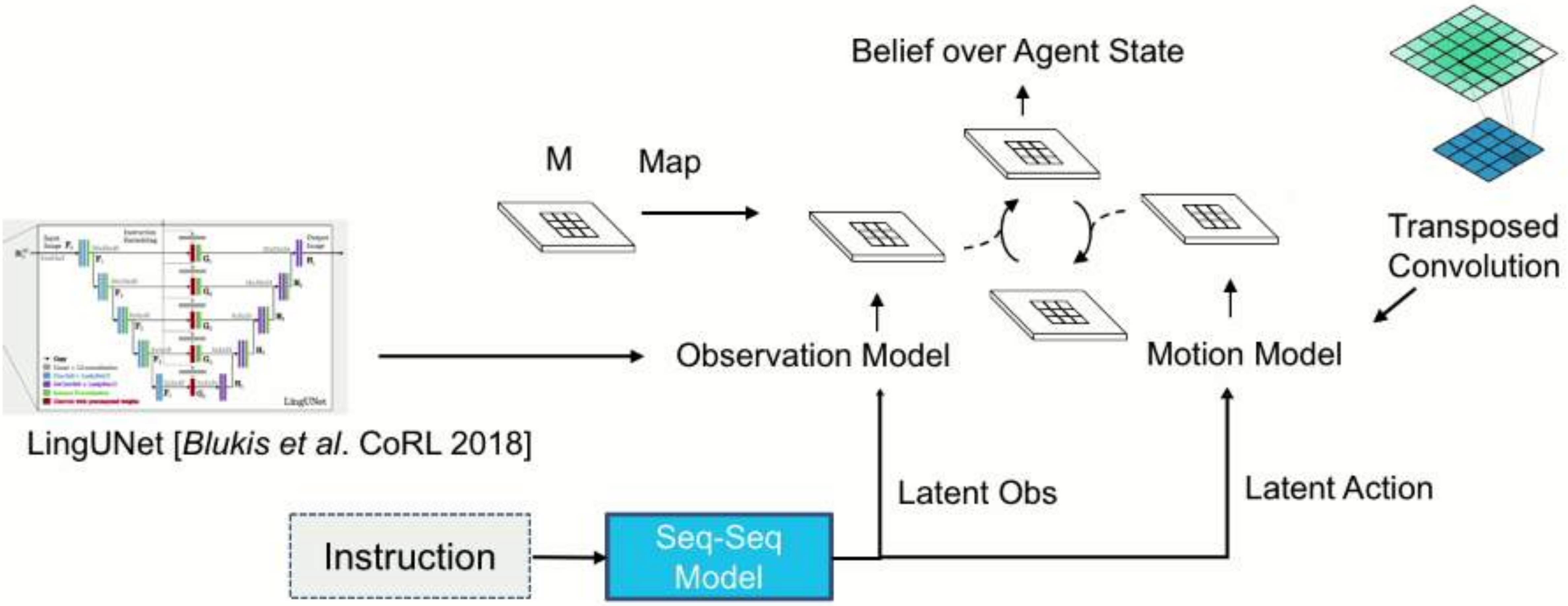
VLN + Bayes Filter

- Incorporates geometric priors via camera projection



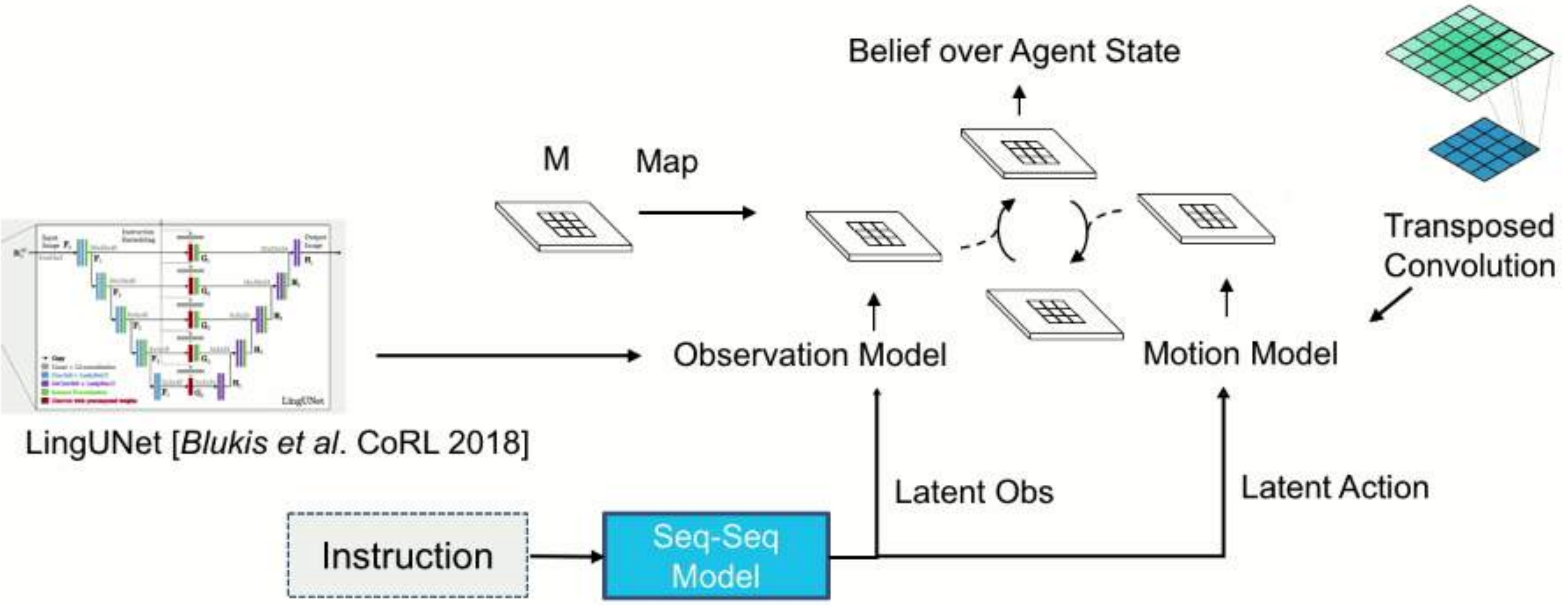
VLN + Bayes Filter

- Incorporates geometric priors via camera projection
- Incorporates useful algorithmic priors via differentiable Bayesian filter – reasons naturally over alternative trajectories, avoids beam search through the environment



VLN + Bayes Filter

- Incorporates geometric priors via camera projection
- Incorporates useful algorithmic priors via differentiable Bayesian filter – reasons naturally over alternative trajectories, avoids beam search through the environment
- Full results on VLN task to come!



Outline

Generating Visually-Grounded Language
(Image Captioning – Novel Object Captioning)

Understanding Visually-Grounded Language
(Vision-and-Language Navigation)

Future Work

Vision and Language

Goal: AI systems that:

- Communicate naturally with people
- Understand visual context

Example: Personal voice-assistants

Image source: TechHive

Image source: Lenovo

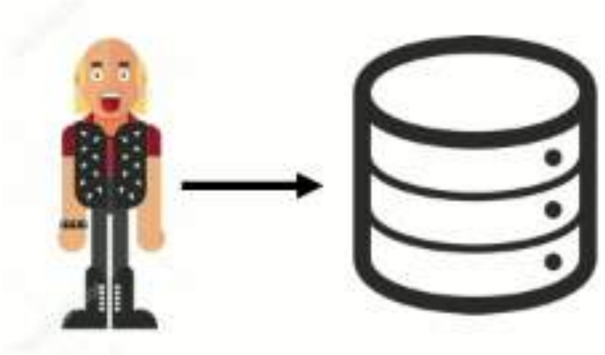
Personalizing AI Systems

Imagine
Use-Cases

Personalizing AI Systems

Imagine
Use-Cases

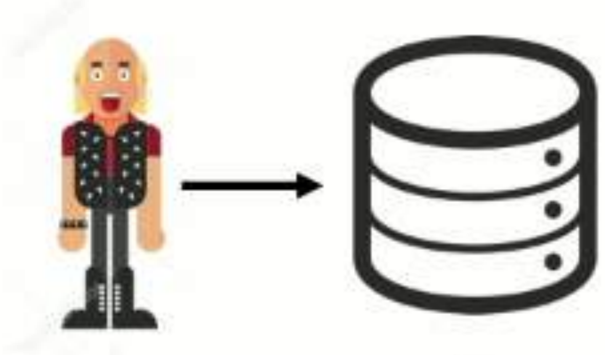
Manufacture
Data



Personalizing AI Systems

Imagine
Use-Cases

Manufacture
Data

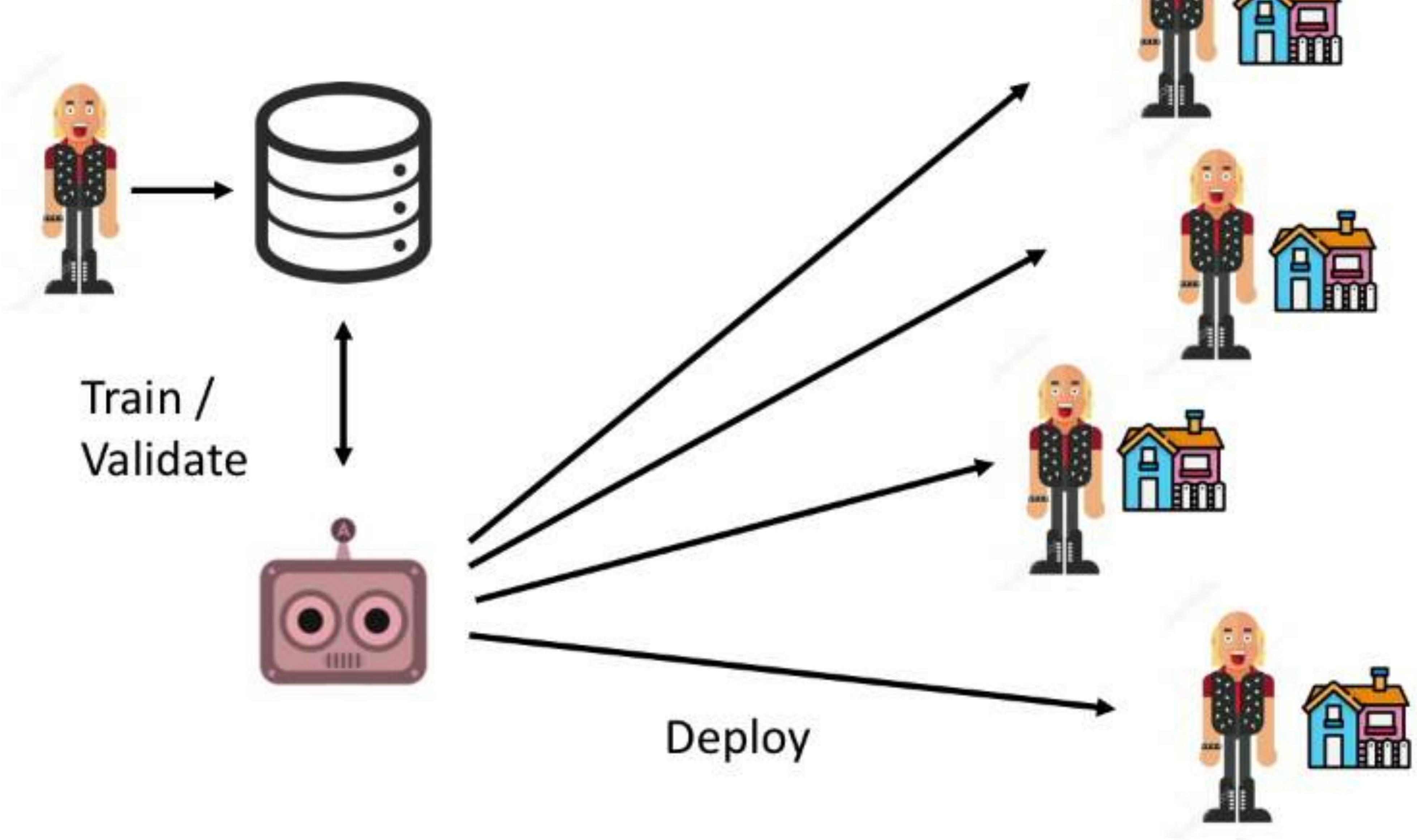


Train /
Validate

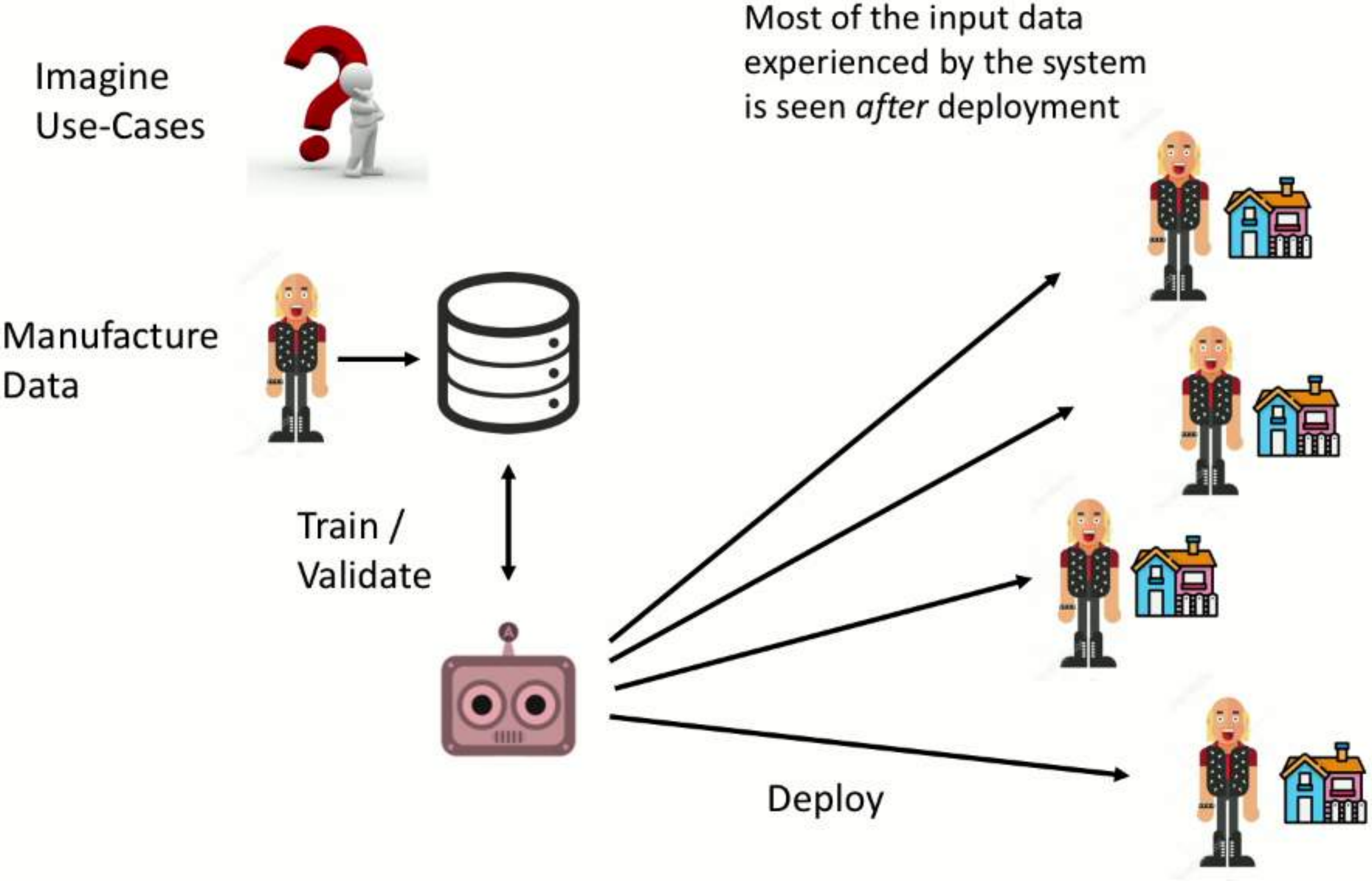
Personalizing AI Systems

Imagine
Use-Cases

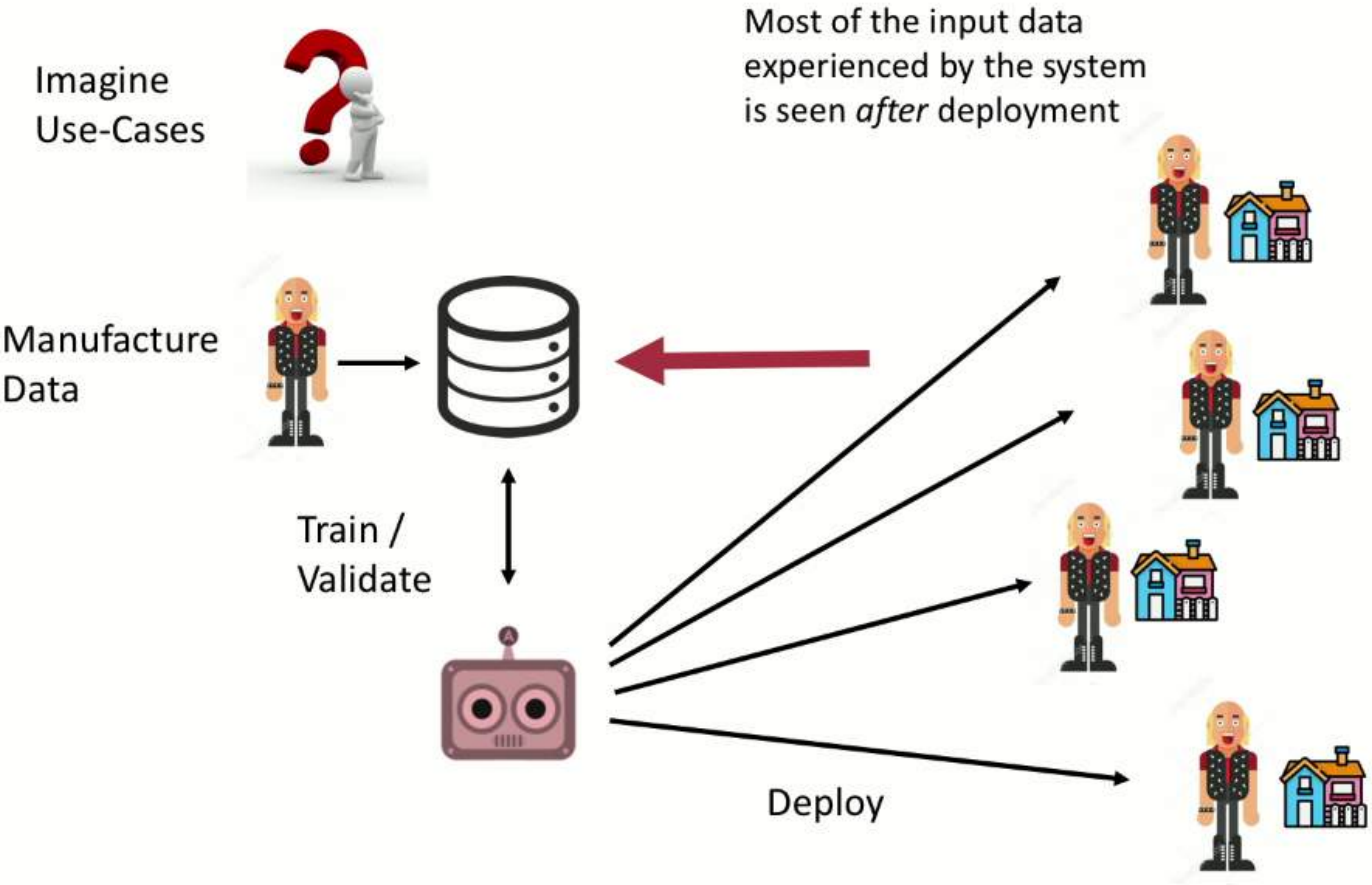
Manufacture
Data



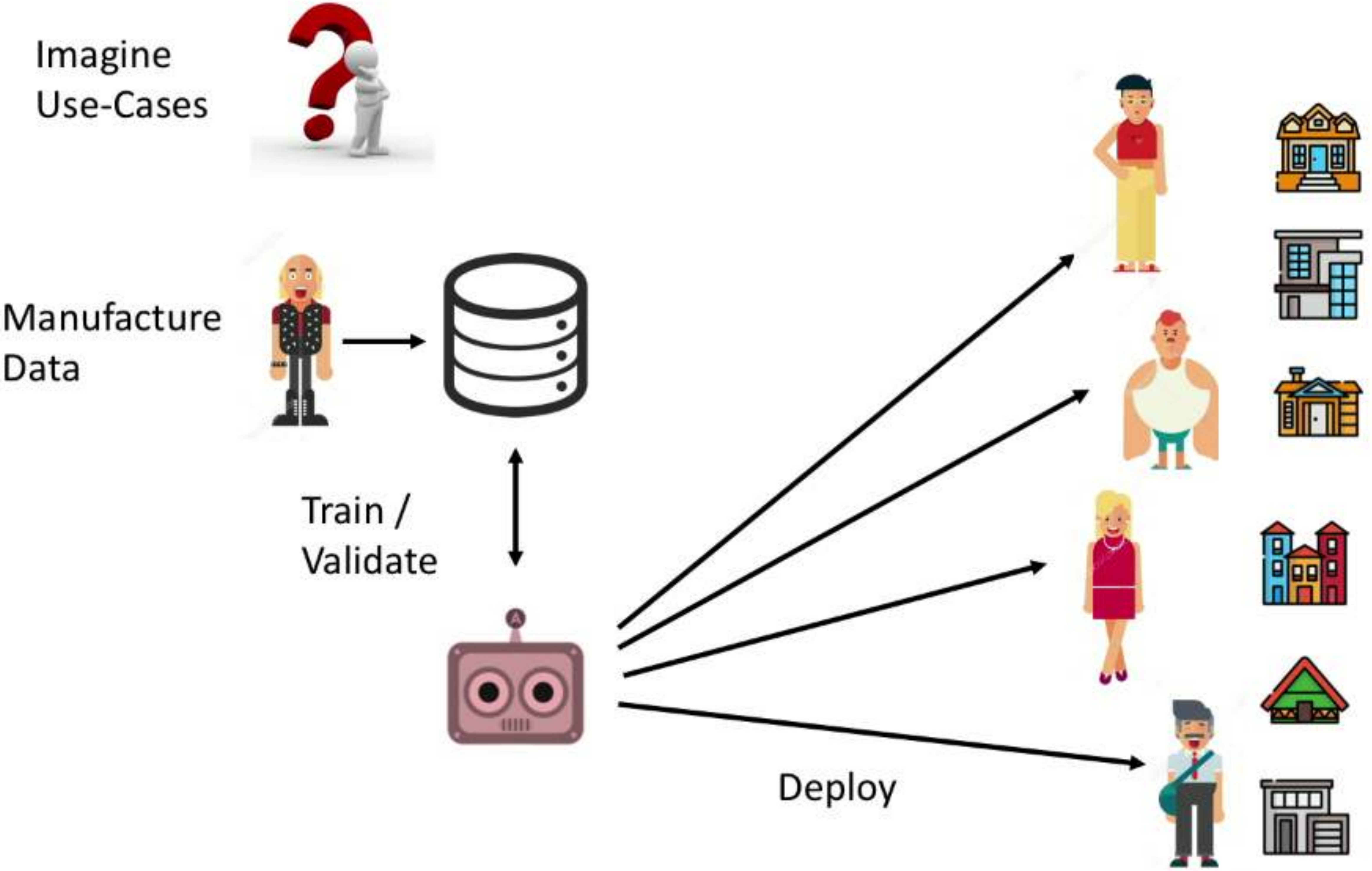
Personalizing AI Systems



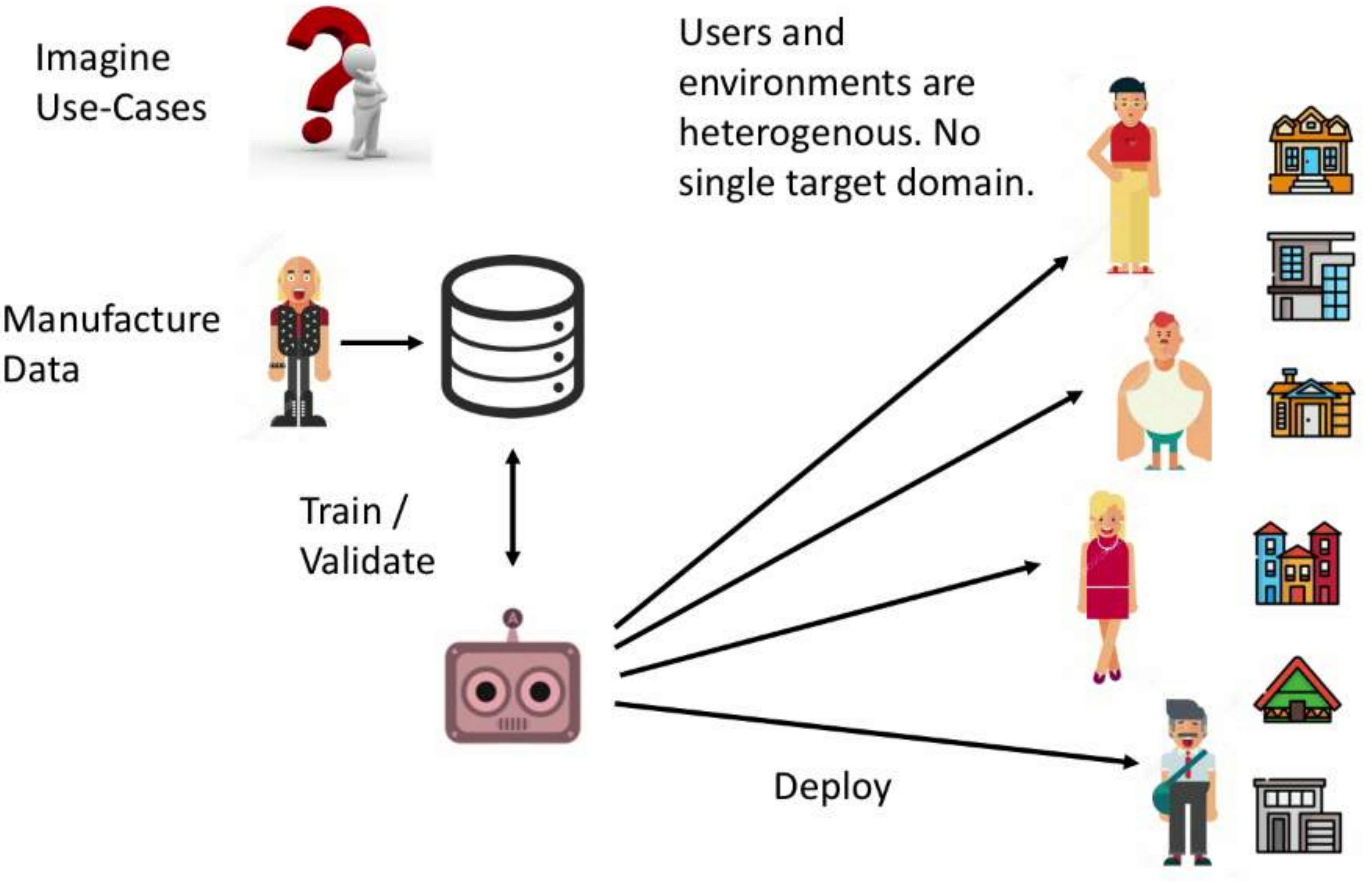
Personalizing AI Systems



Personalizing AI Systems



Personalizing AI Systems



Personalizing AI Systems

- Solution – put more power in the hands of the user
- It's okay – users are smart and creative!

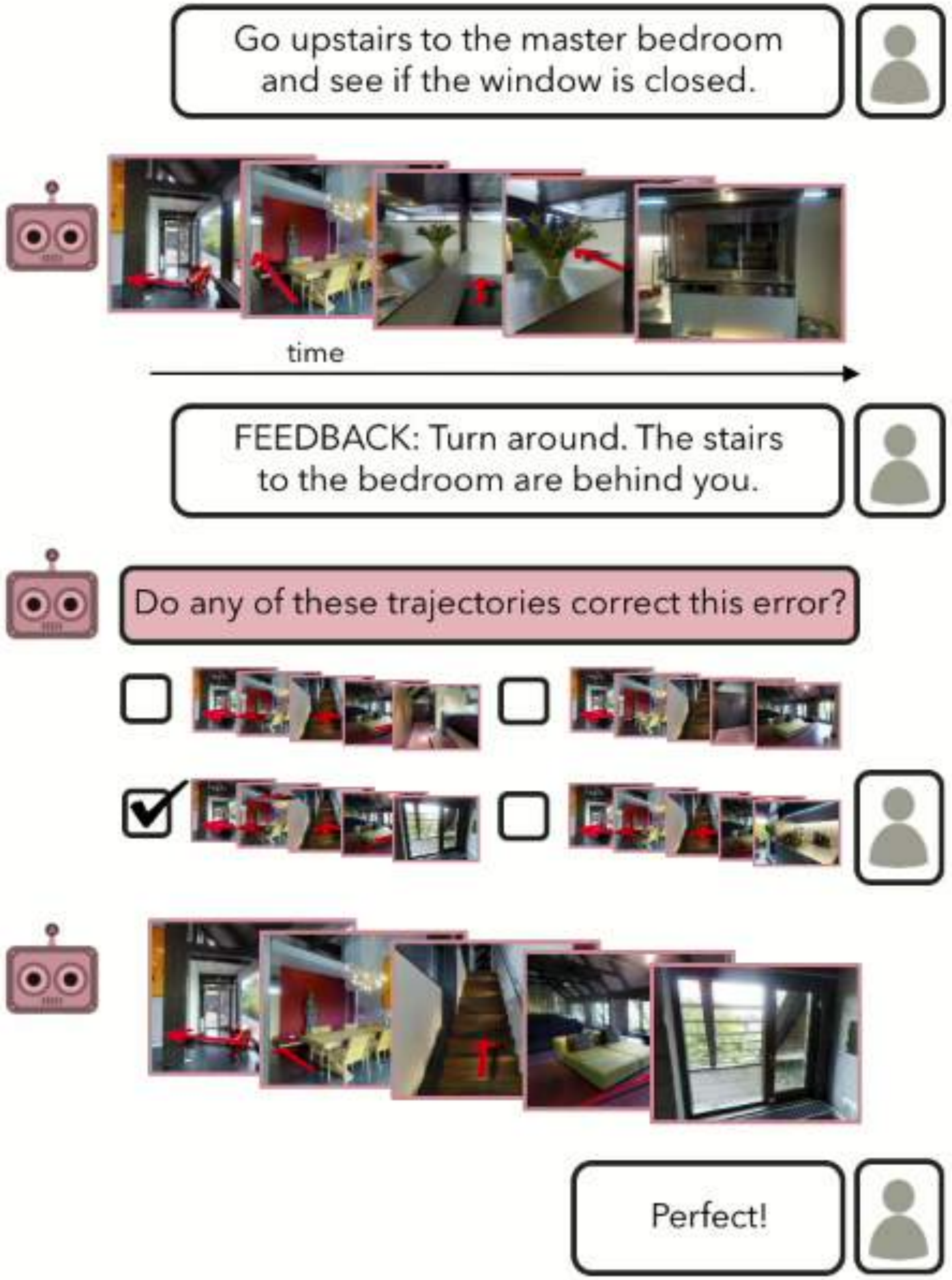
Planetminecraft.com

Hacking Roomba, 2007

Ultimate Simpsons Doom mod

Personalizing AI Systems

Personalizing AI Systems



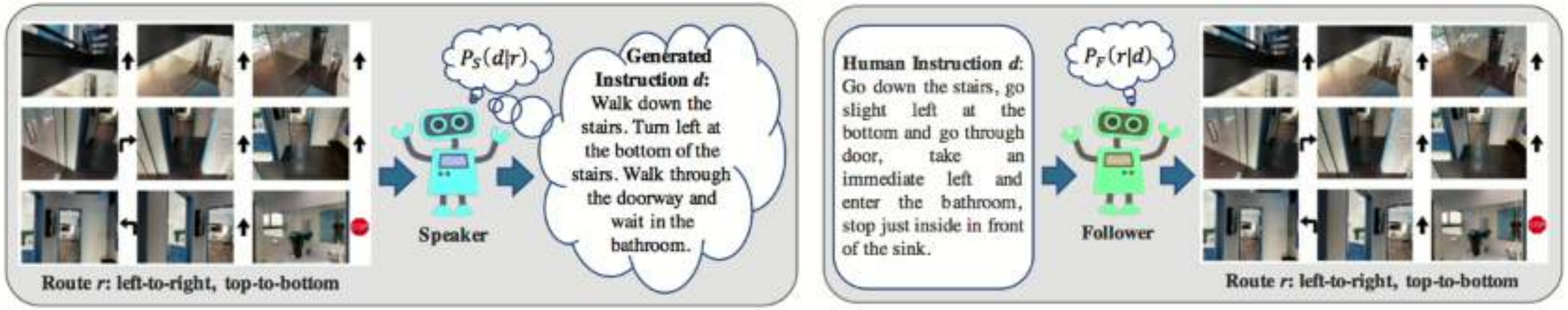
Challenges:

- Generating and representing agent intent
- Natural language feedback
- Large-scale meta-learning for personalization

Cooperative Dialog Training

Great success from:

- Data augmentation in computer vision
- Back Translation in machine translation
- Self-play in RL, e.g. Dota, Go
- But - cooperative training of dialog agents still a major challenge!
- Algorithmic priors may be part of the solution



[*Fried et al. NeurIPS* 2018]

Acknowledgments

Rishabh Jain
Georgia Tech

Xinlei Chen
FAIR

Mark Johnson
Macquarie University

Yufei Wang
Macquarie University

Ayush Shrivastava
Georgia Tech

Harsh Agrawal
Georgia Tech

Karan Desai
Georgia Tech

Stefan Lee
Georgia Tech

Devi Parikh
Georgia Tech / FAIR

Dhruv Batra
Georgia Tech / FAIR

- Stephen Gould, Australian National University
- Basura Fernando, Australian National University
- Anton van den Hengel, University of Adelaide
- Ian Reid, University of Adelaide
- Qi Wu, University of Adelaide
- Damien Teney, University of Adelaide
- Jake Bruce , Queensland University of Technology
- Niko Sünderhauf, Queensland University of Technology
- Xiaodong He, Microsoft Research
- Lei Zhang, Microsoft Research
- Chris Buehler, Microsoft Research

Thanks!

Questions?