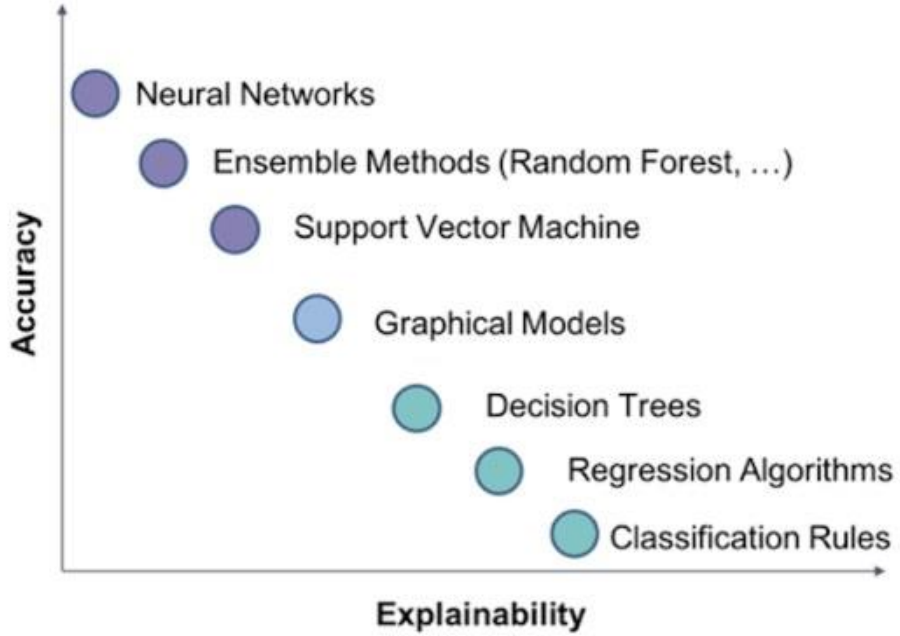
Recent progress towards XAI at UC Berkeley

XAI Workshop, ICCV 2019

Prof. Trevor Darrell

Inverting the Accuracy-Explainability Curve

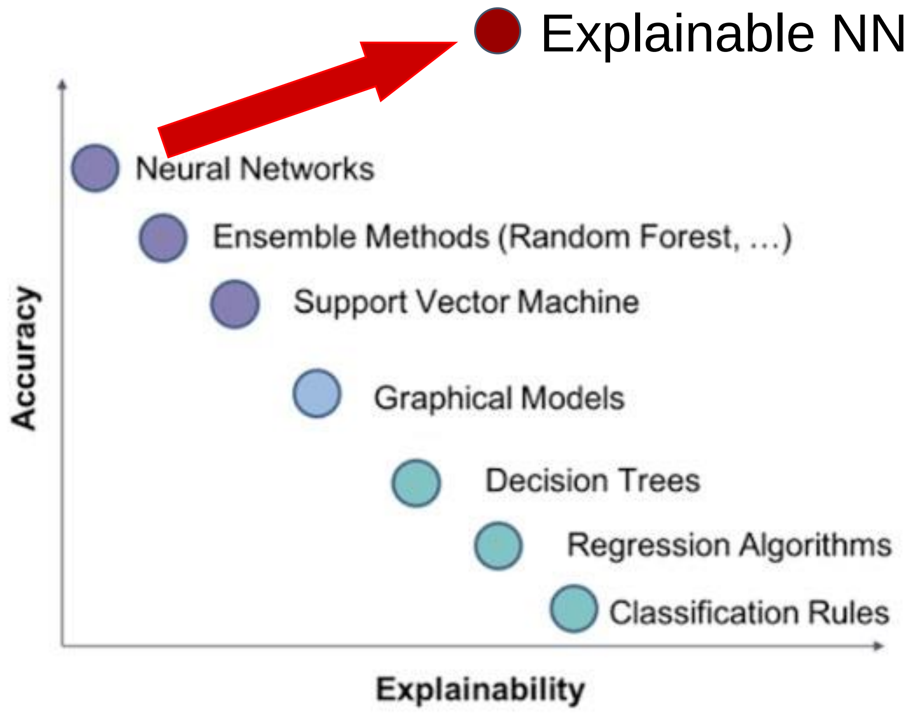
- The usual XAI story:



Inverting the Accuracy-Explainability Curve

- Despite conventional wisdom, adding explainability to deep AI models does not decrease their accuracy, and can even improve it.

- We don't need to choose between explainability or high accuracy, can have both!



Inverting the Accuracy-Explainability Curve

DNN XAI systems can lead to *better-performing, more explainable* models:

1. **Explanations-as-additional-loss**: adding the “show your work” and “right for the right reasons” constraint.

2. **Explanations-allow-advice:** XAI systems transduce DNN states to natural language; reversing this, we can create “Advisable AI” and refine a model via language guidance rather than additional labeled examples.

3. **Explanations-reveal-model-uncertainty:** in a human-in-the-loop retrieval system, explanations let human operators more accurately judge when they should accept suggestions from an XAI teammate.

Today

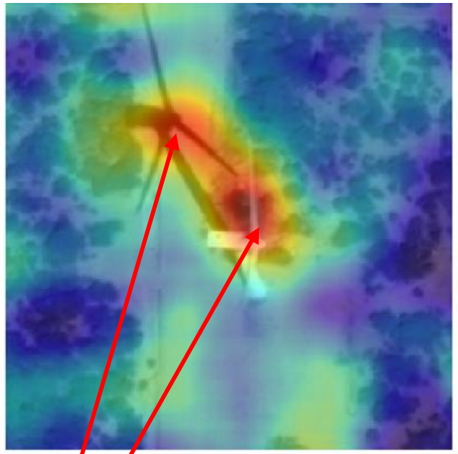
- Multi-step Saliency via Compositional NMNs
- Fine-grained Textual Explanations
- From Explainable to “Advisable”Driving Models

Attentive XAIs *do not decrease* recognition accuracy

RISE[1] XAI provides saliency explanations to AI models without affecting their accuracy.

Wind farm: 100%

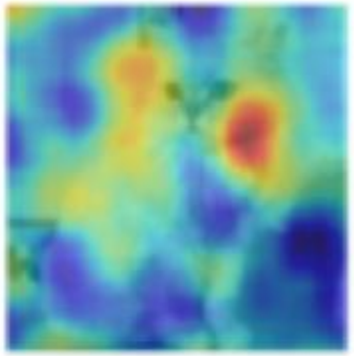
Explanation for "wind farm"



most important (red)

Why does the AI system think these two photos are similar?

because ...



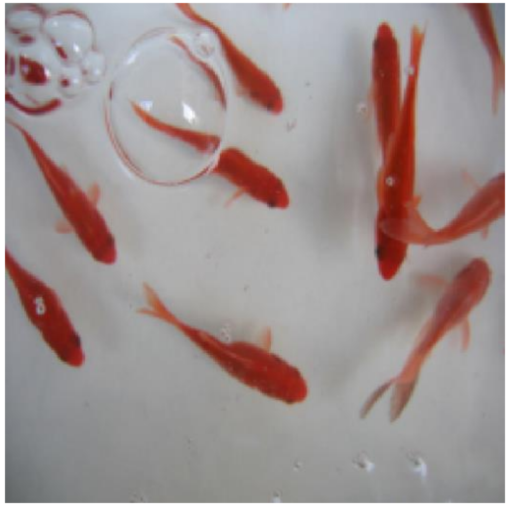
"furry"

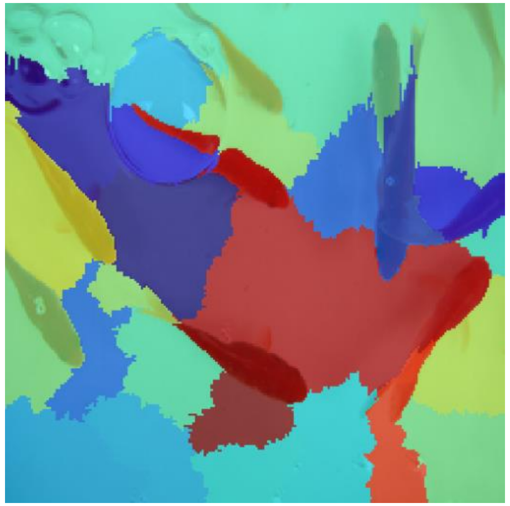
SANE[2]: When enforcing explainability, attribute recognition performance **improves** by 2-3% mAP on two diverse datasets.

[1] Vitali Petsiuk, Abir Das, Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. BMVC Oral, 2018
[2] Plummer et al. Why do These Match? Explaining the Behavior of Image Similarity Models, 2019

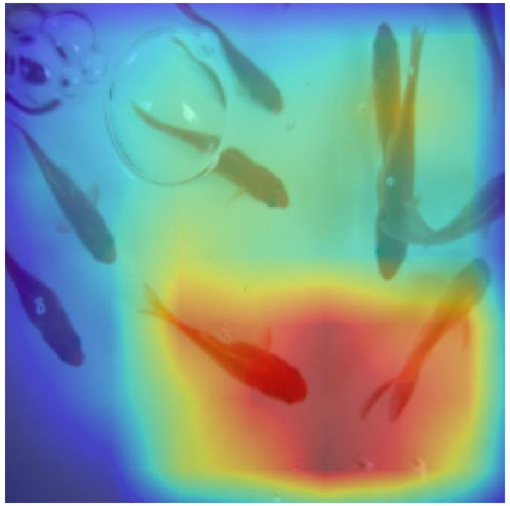
Salience for Introspection

Goldfish

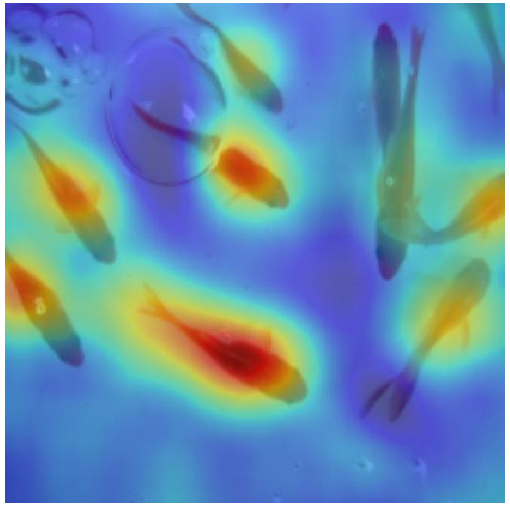




LIME



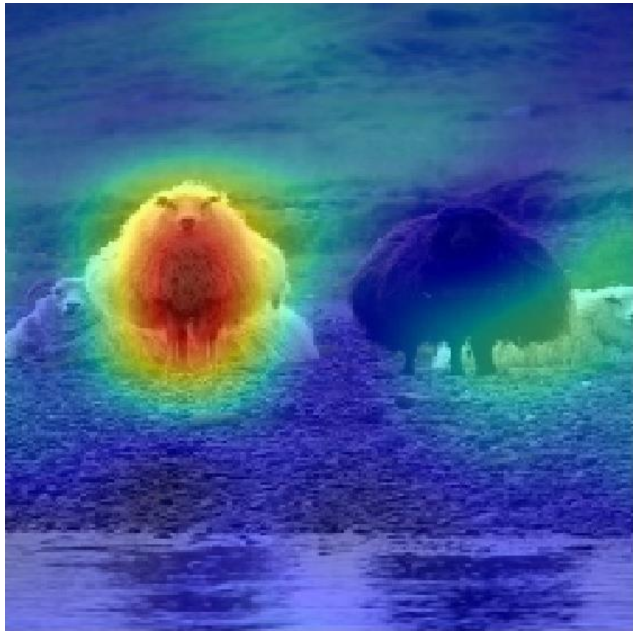
GradCAM



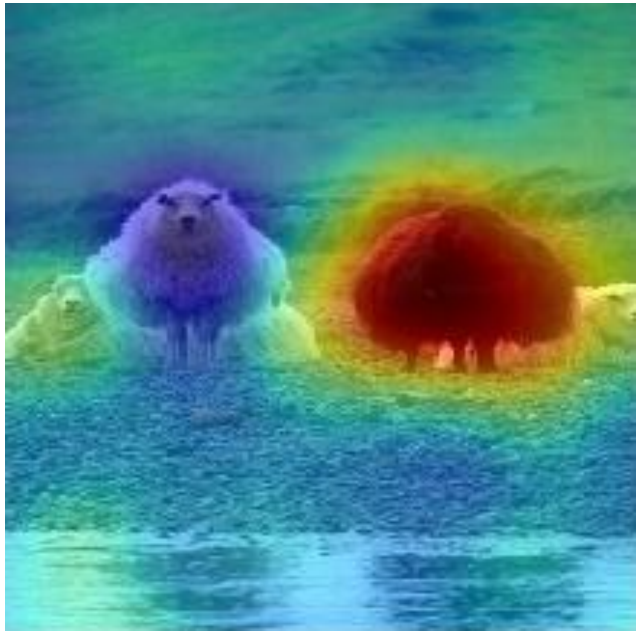
RISE

- RISE probes black-box CNN models with randomly masked instances of an image to find class-specific evidence

RISE can explain different categories



Explanation for **Sheep**



Explanation for **Cow**

RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018

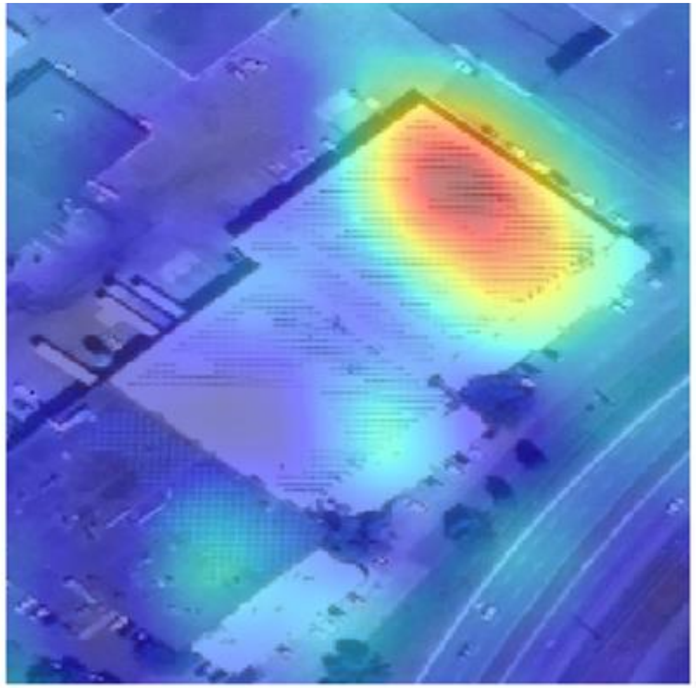
RISE: Randomized Input Sampling for Explanation

Neural network prediction:

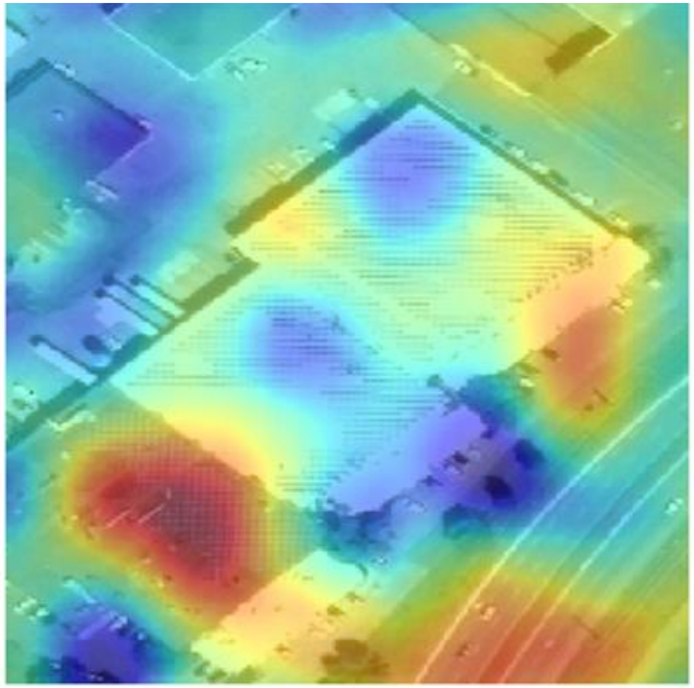
solar farm: 63%, **shopping mall**: 23%

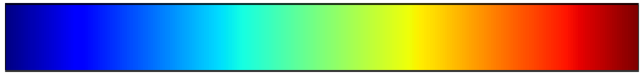
Image from the FMoW dataset

RISE
Explanation for **solar farm**



RISE
Explanation for **shopping mall**

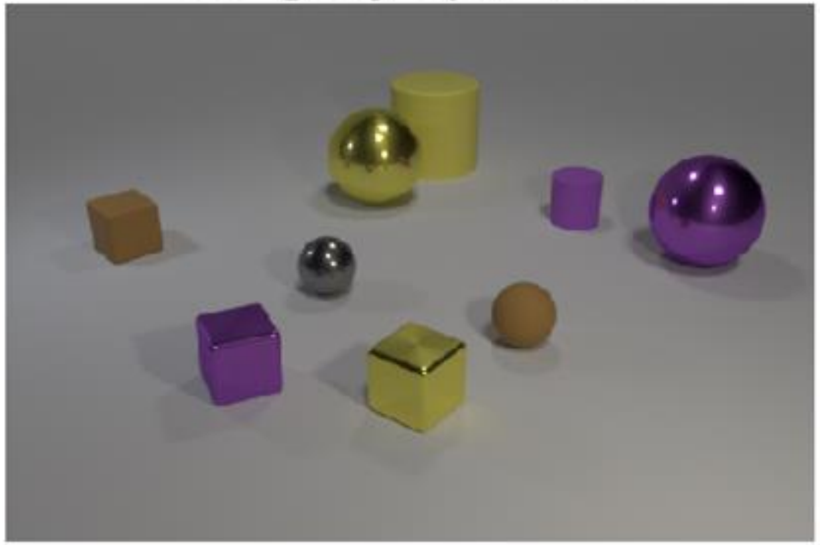


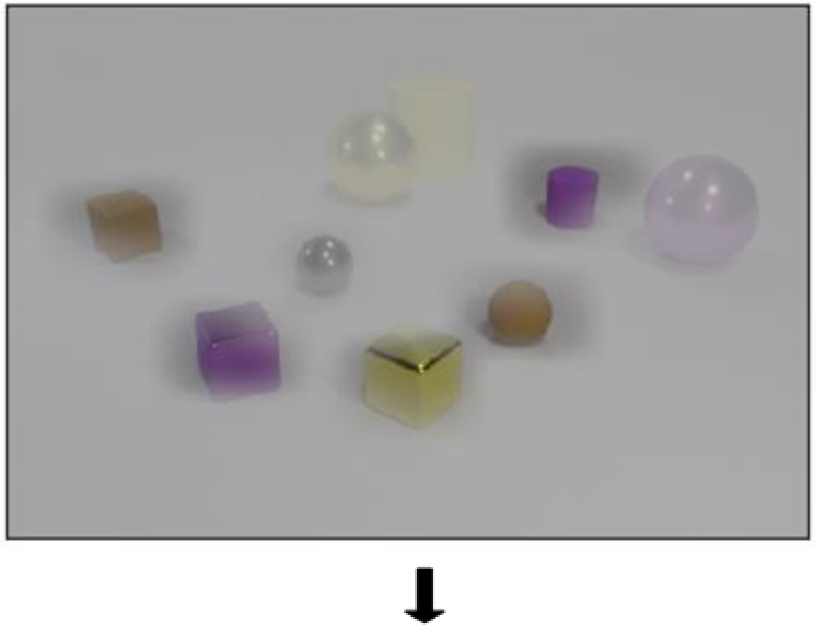


Increasing importance

Single-step Salience is limited

what number of other objects are
there of the same size as
the gray sphere ?

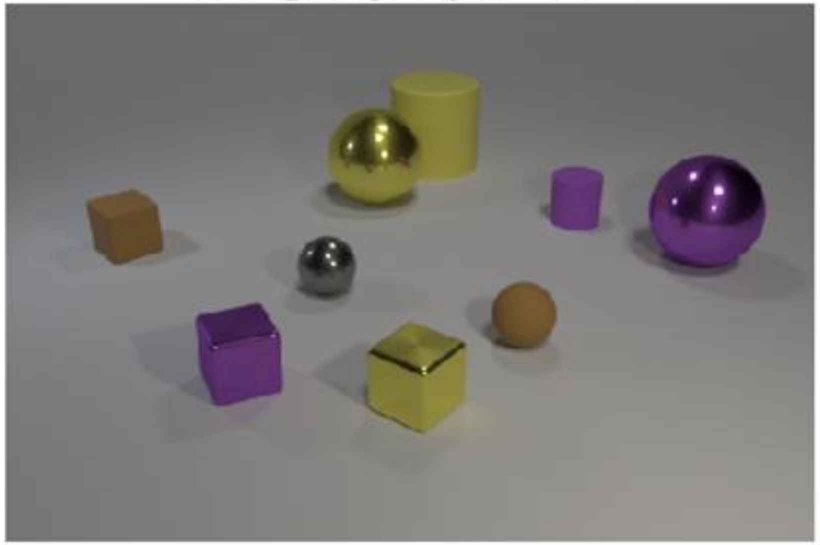




predicted answer: "5"

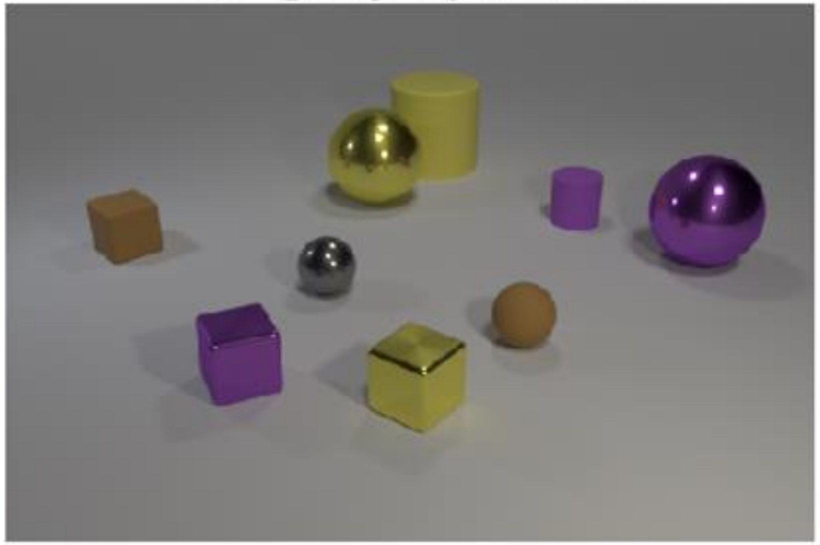
Multi-step introspection / transparent reasoning

what number of other objects are there of the same size as the gray sphere ?



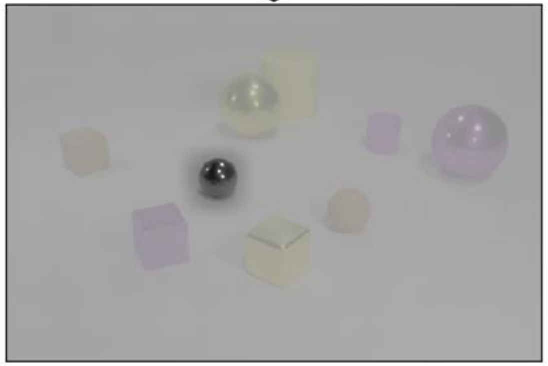
Multi-step introspection / transparent reasoning

what number of other objects are
there of the same size as
the gray sphere ?



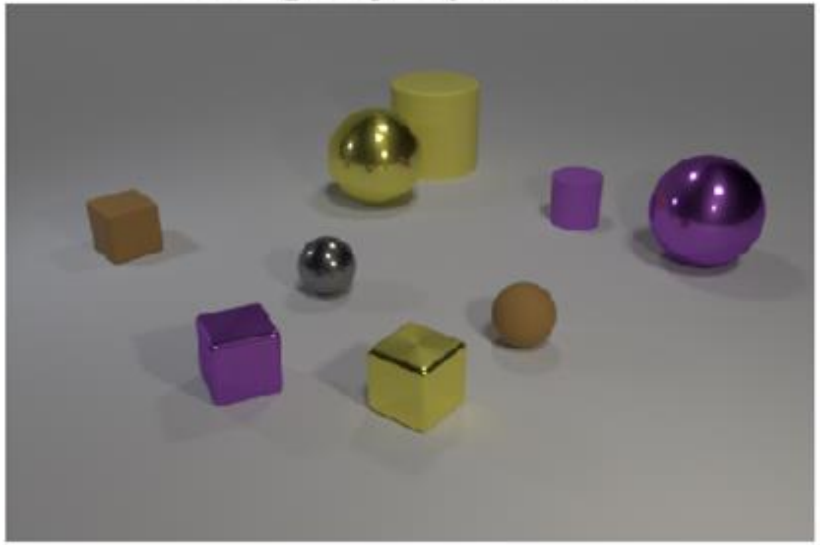
Reasoning
Step 1

look_for("gray sphere")



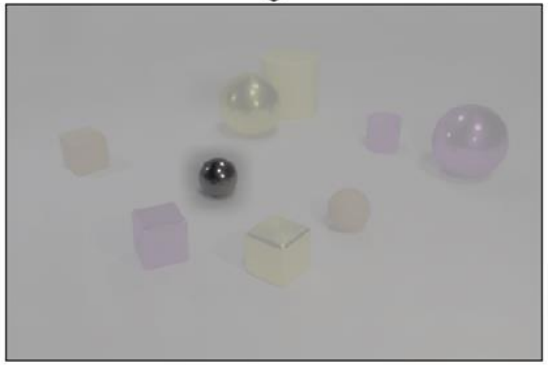
Multi-step introspection / transparent reasoning

what number of other objects are
there of the same size as
the gray sphere ?



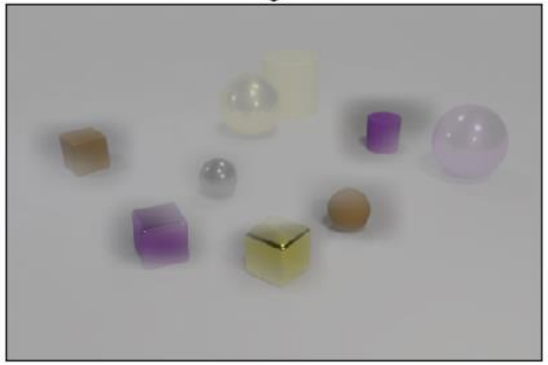
Reasoning
Step 1

look_for("gray sphere")



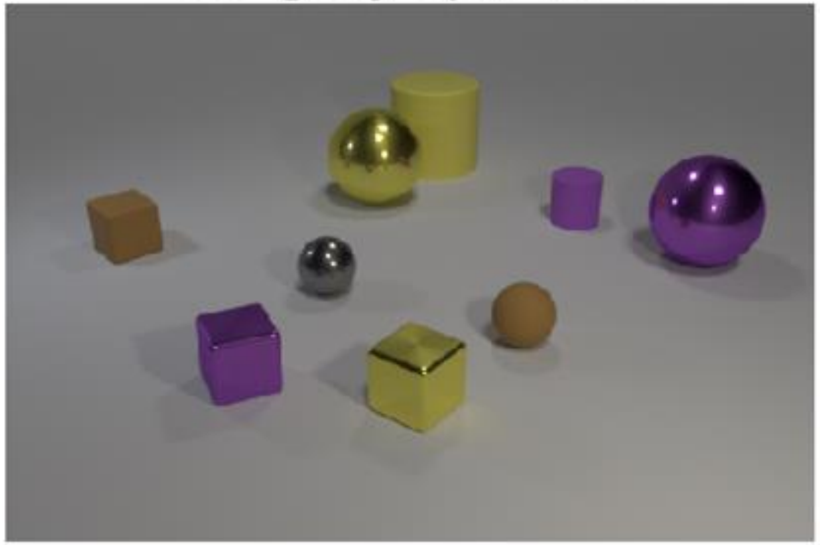
Reasoning
Step 2

related_by("size")



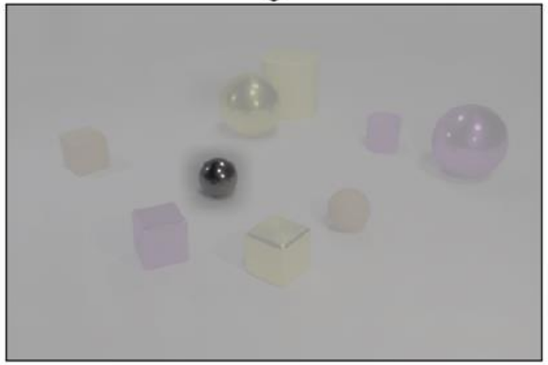
Multi-step introspection / transparent reasoning

what number of other objects are there of the same size as the gray sphere ?



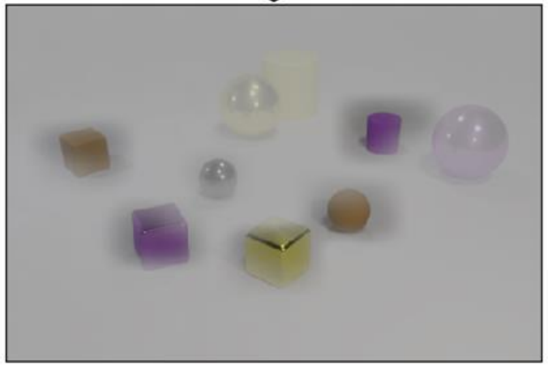
look_for("gray sphere")

Reasoning **Step 1**



related_by("size")

Reasoning **Step 2**



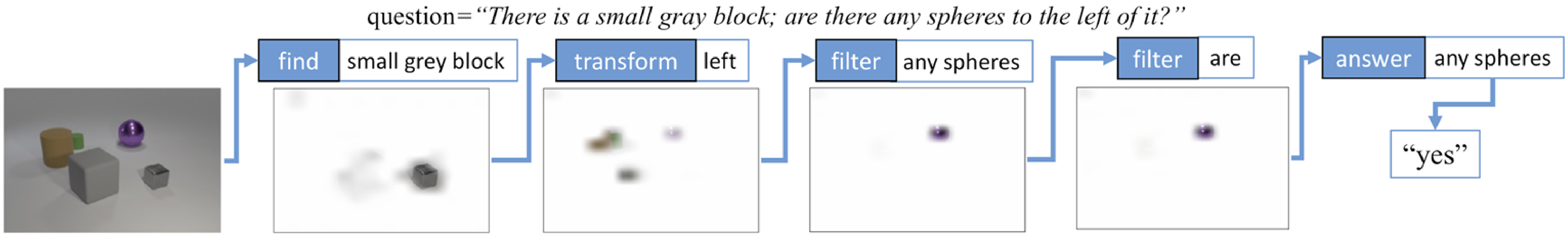
Reasoning **Step 3**

answer("number", "other objects")

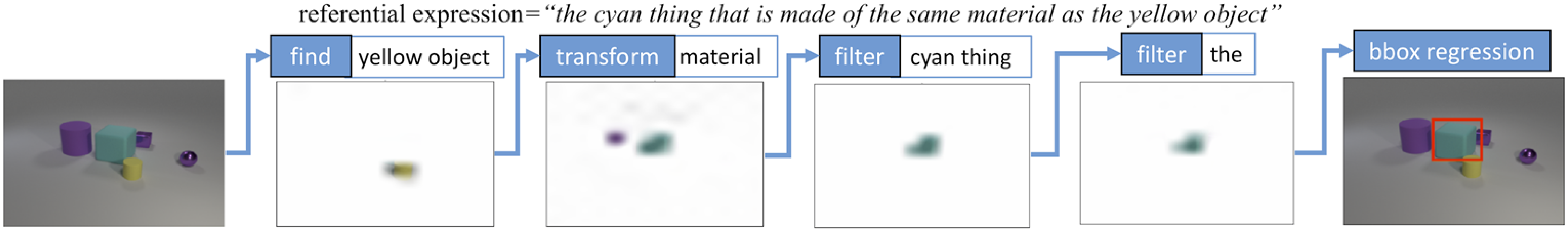
predicted answer: "5"
true answer: "5"

Neural module networks

Example predictions on Visual Question Answering (VQA)



Example predictions on Referential Expression Grounding (REF)



Monolithic Networks for Visual Question Answering

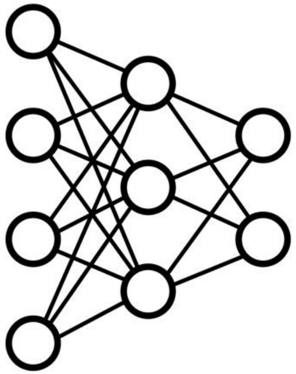
What is this?

a cat

Monolithic Networks for Visual Question Answering

Monolithic Networks

✓ Work well on simple questions



a cat

What is this?

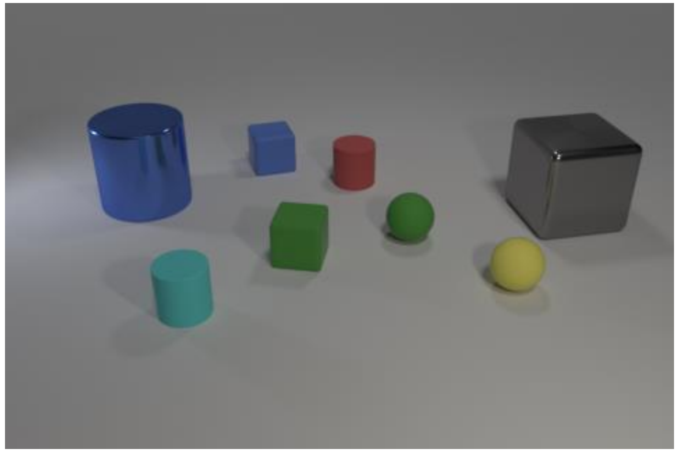
Monolithic Networks for Visual Question Answering

Monolithic Networks

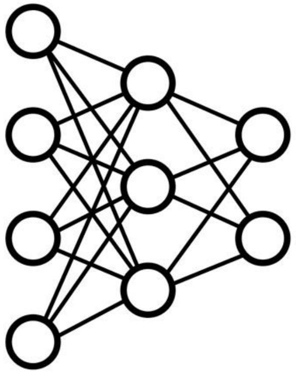
✓ Work well on simple questions

✗ Challenging for questions requiring *compositional reasoning*

✗ Limited interpretability



What color is the thing with the same size as the blue cylinder?

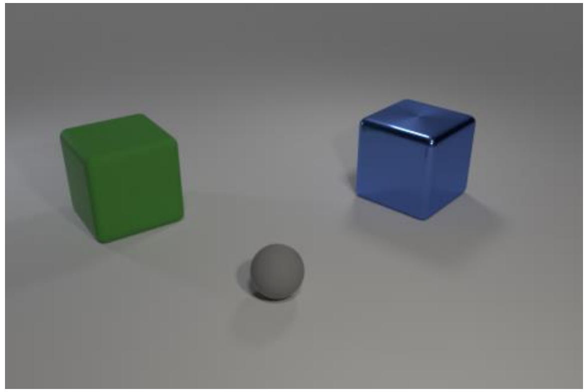


?

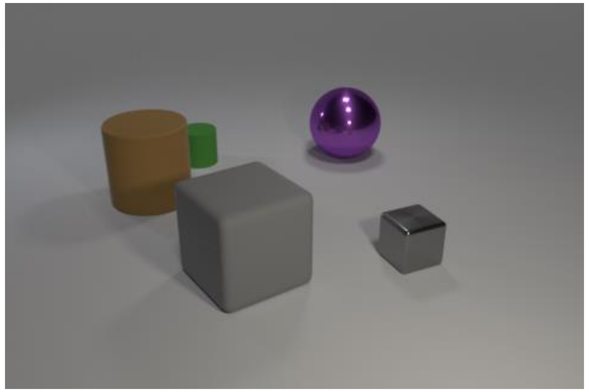
Compositionality in Reasoning

- Generalization to complicated **unseen reasoning structure** of **seen operations (relations)**

training time

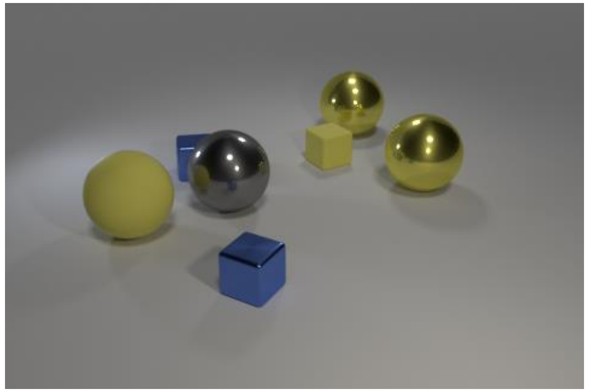


how many objects are the either green rubber object or blue cubes?



is there a big brown object of the **same size** as the green thing?

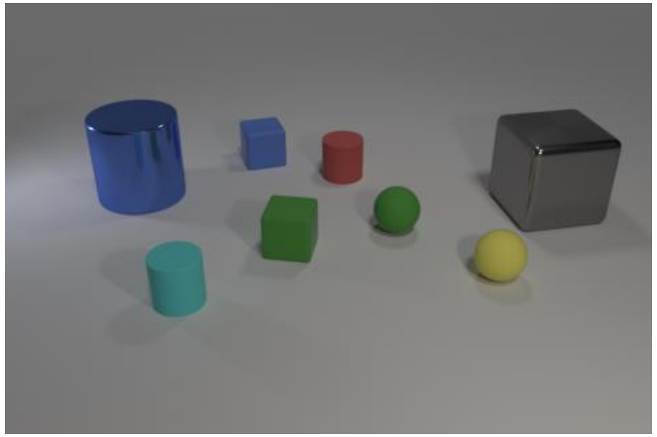
test time



how many other things are the **same size** as the yellow rubber ball?

Compositional Inference with **Modules**

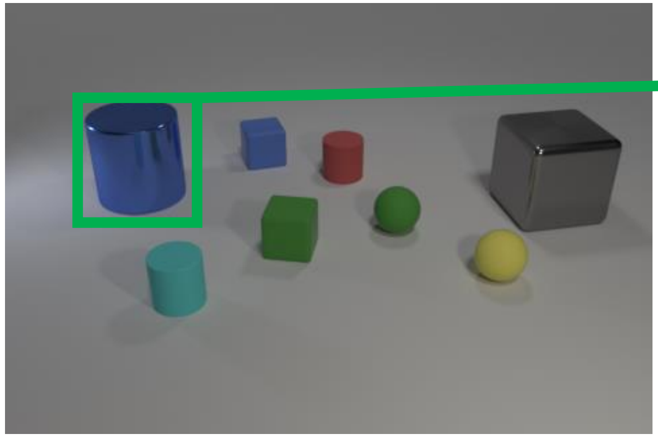
What color is the thing with the same size as the blue cylinder?



```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

Compositional Inference with **Modules**

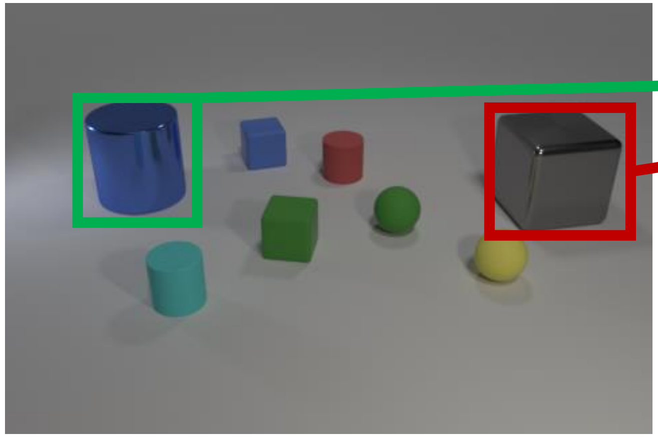
What color is the thing with the same size as the blue cylinder?



```
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

Compositional Inference with **Modules**

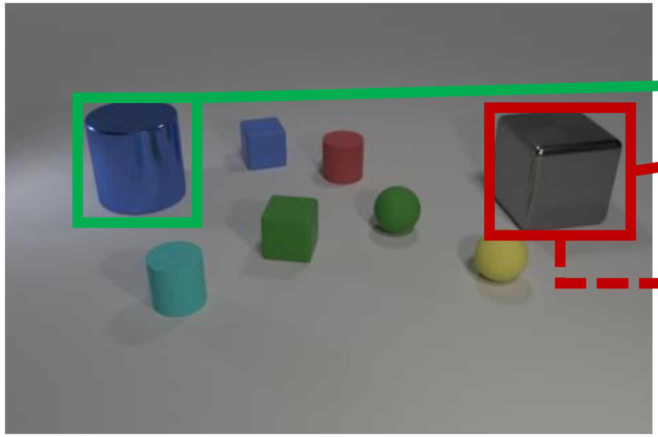
What color is the thing with the same size as the blue cylinder?



```
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

Compositional Inference with **Modules**

What color is the thing with the
same size as the blue cylinder?



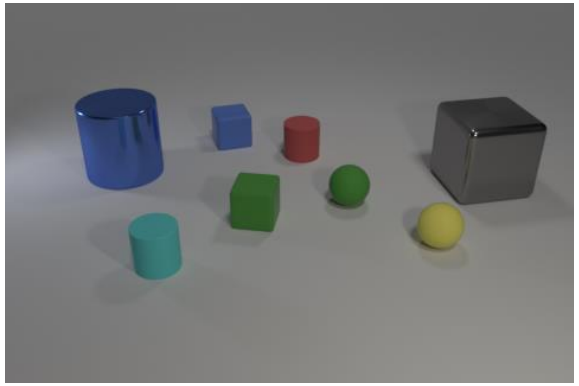
```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

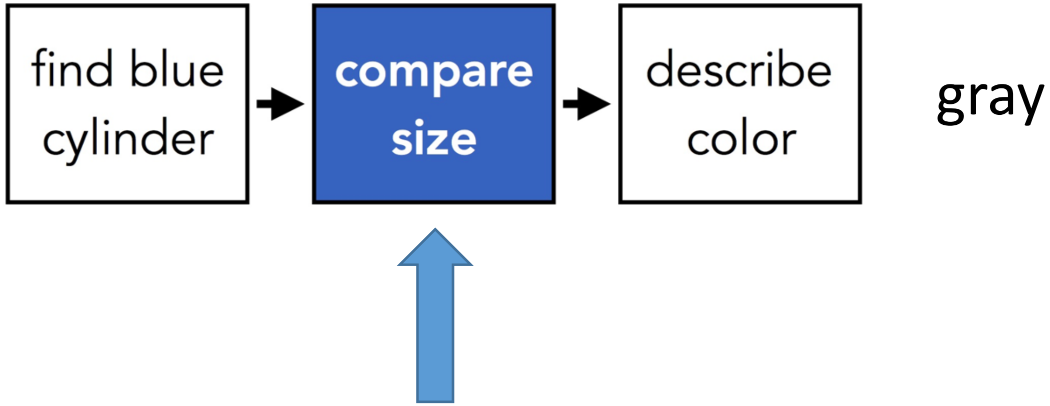
output answer: “gray”

- **Predict** a discrete *execution graph* of modules to answer complex natural language questions

Neural Module Networks (NMNs)

What color is the thing with the same size as the blue cylinder?

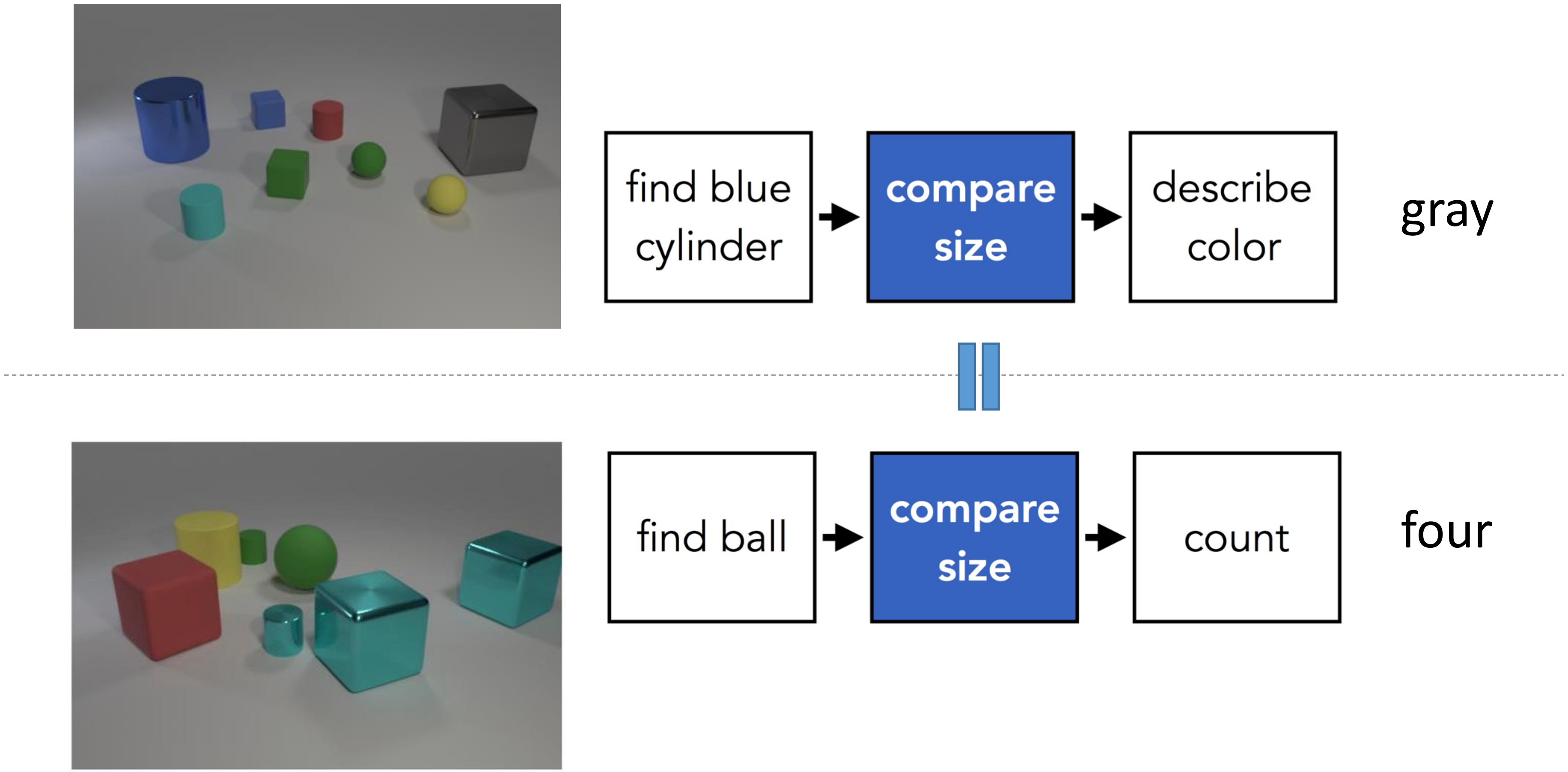





```python
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```

Dynamic and Reusable Modules

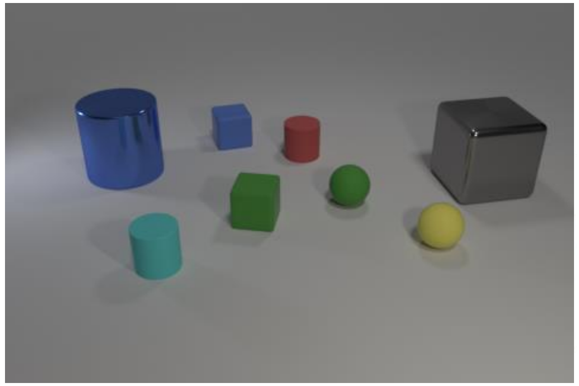
What color is the thing with the same size as the blue cylinder?

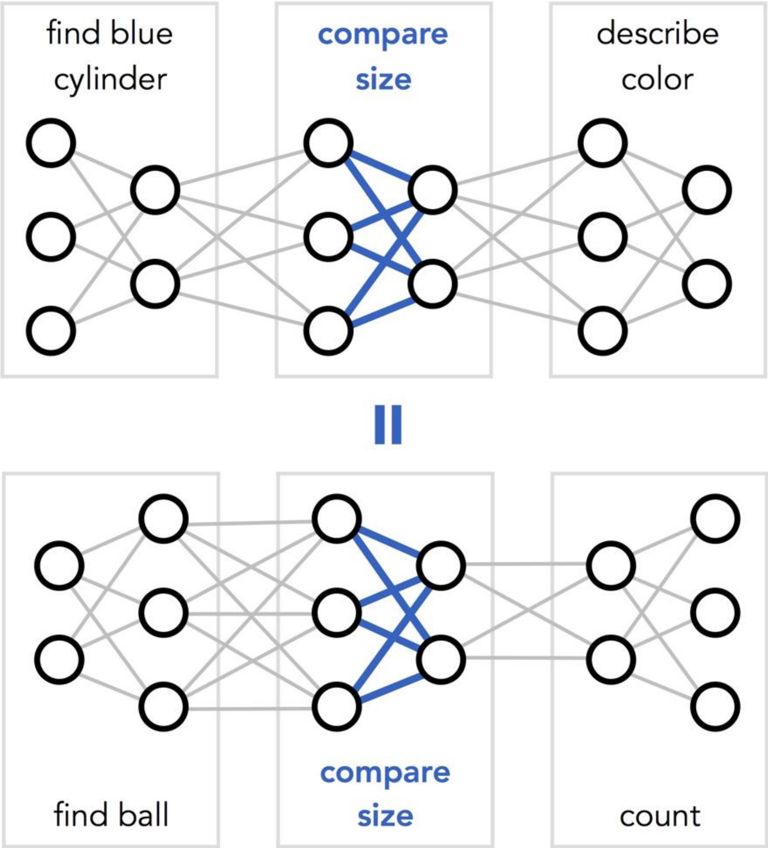


How many things are the same size as the ball?

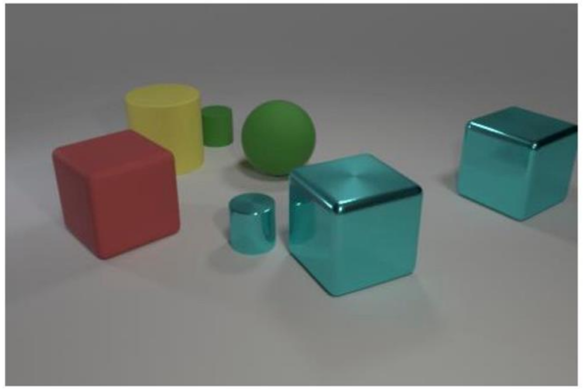
Dynamic and Reusable Modules

What color is the thing with the same size as the blue cylinder?





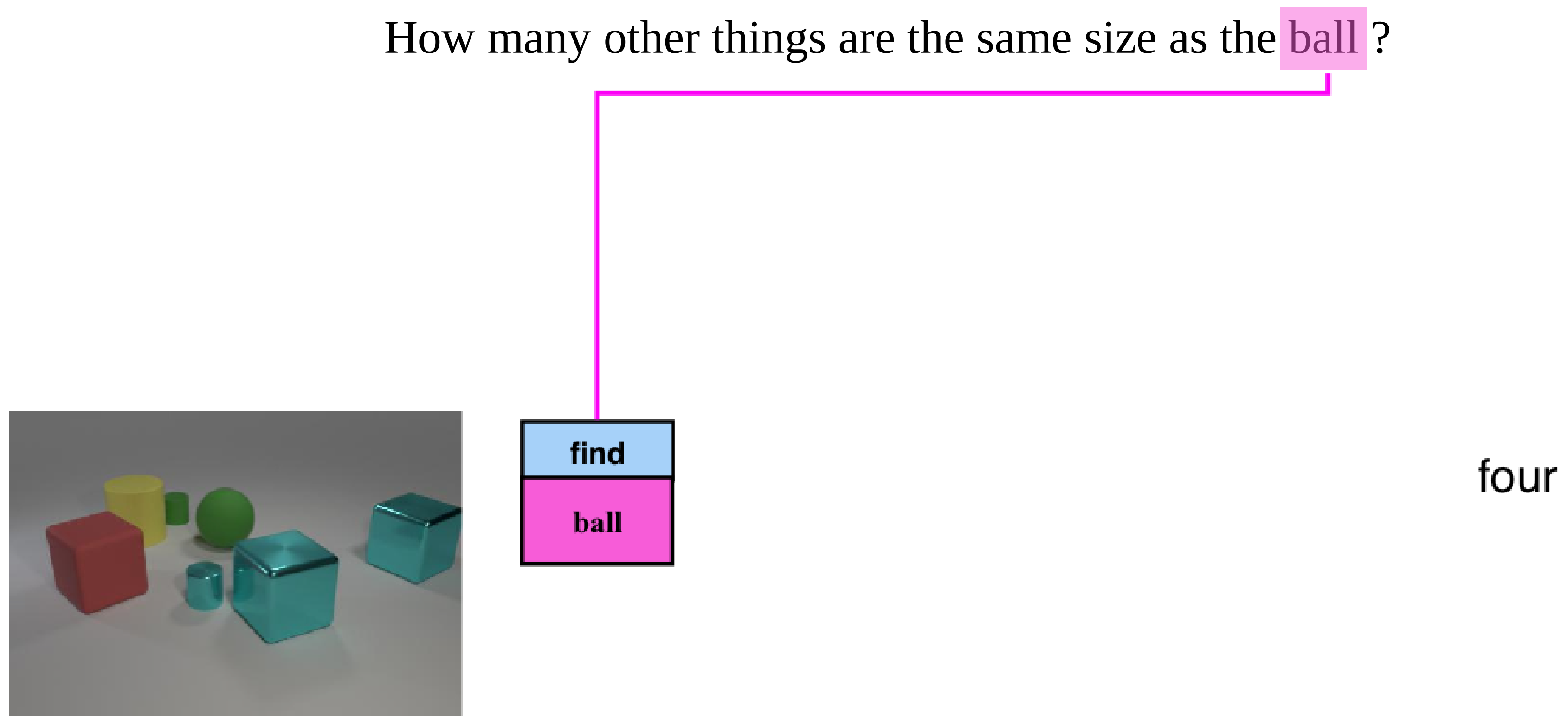
gray



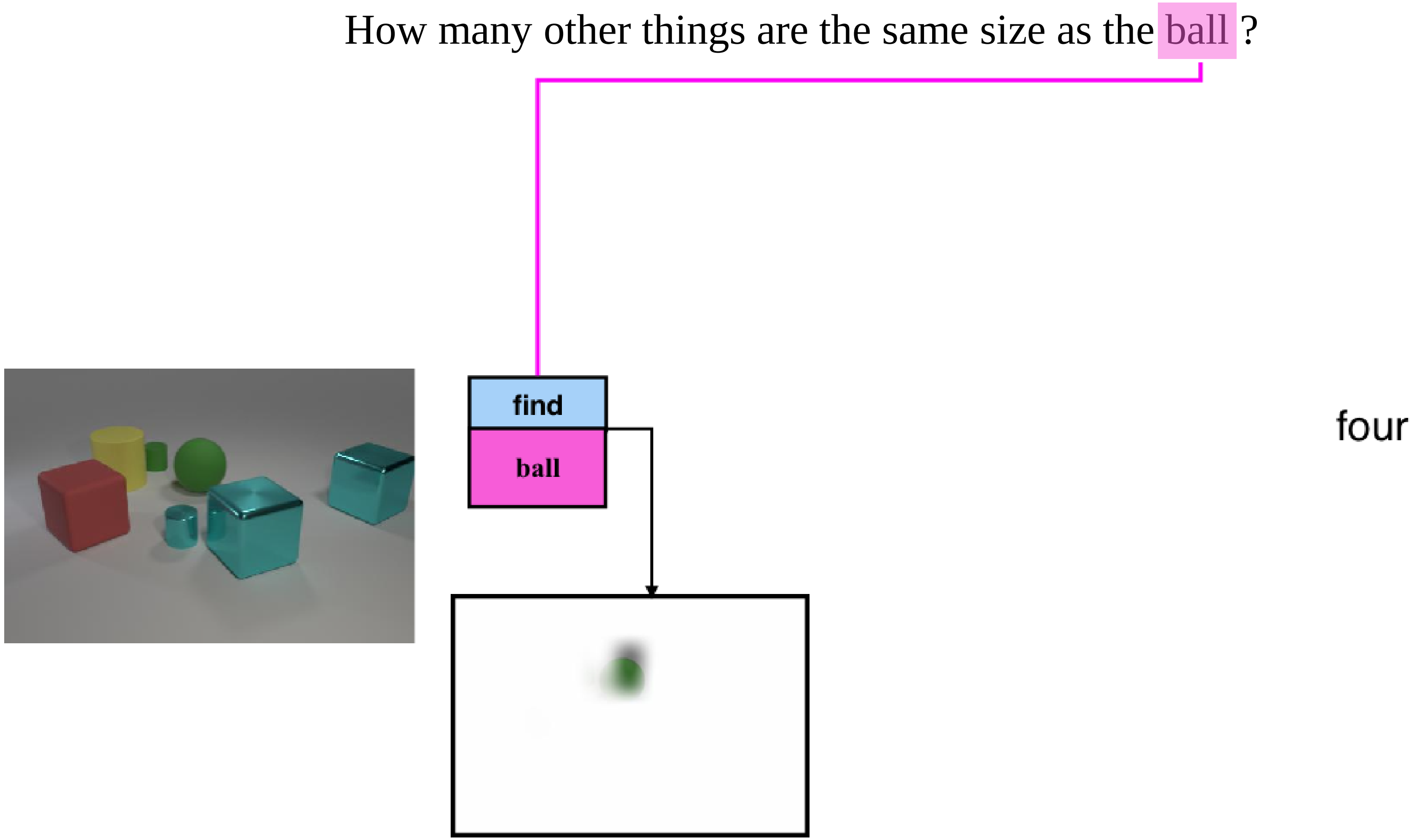
four

How many things are the same size as the ball?

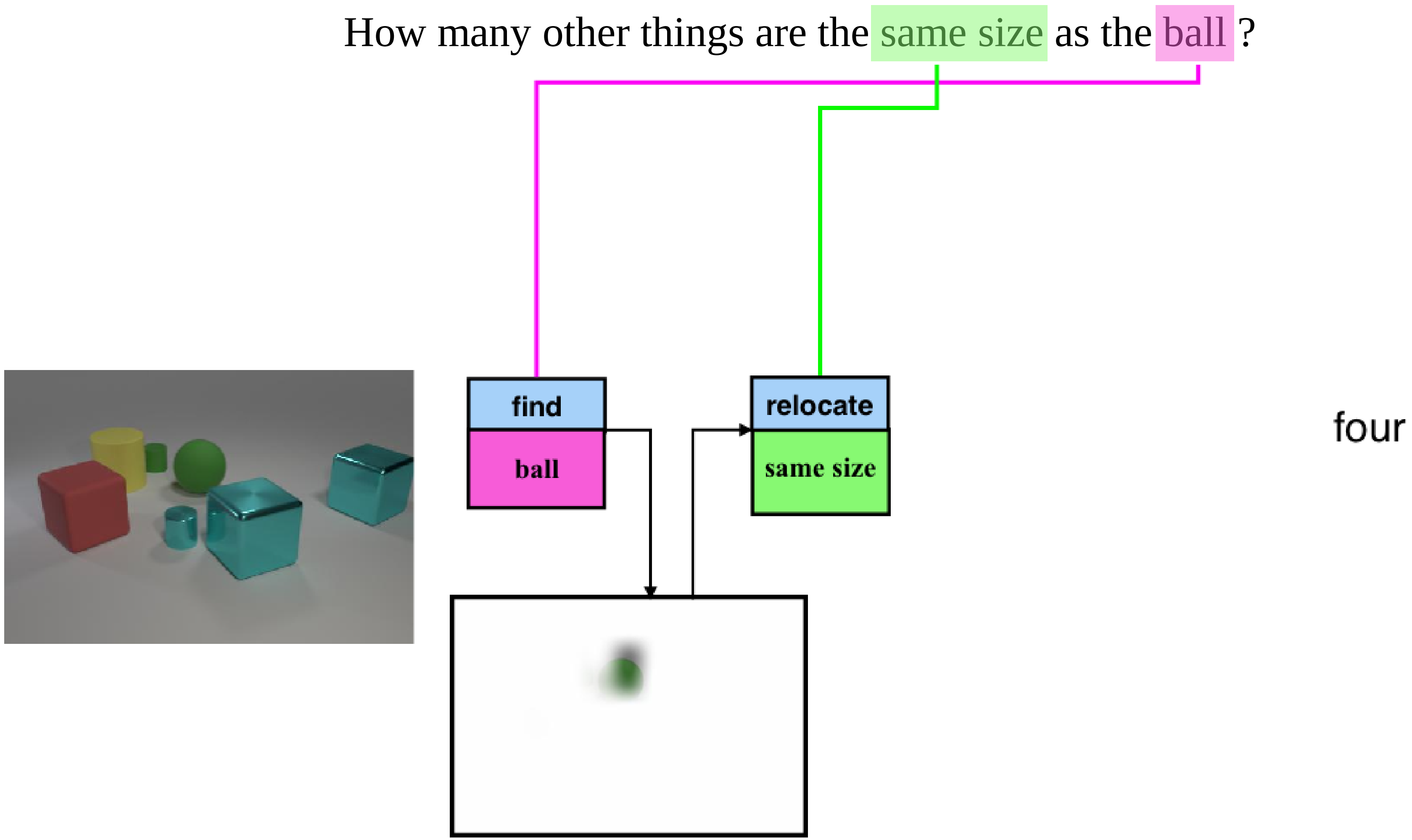
End-to-End Module Networks (N2NMN)



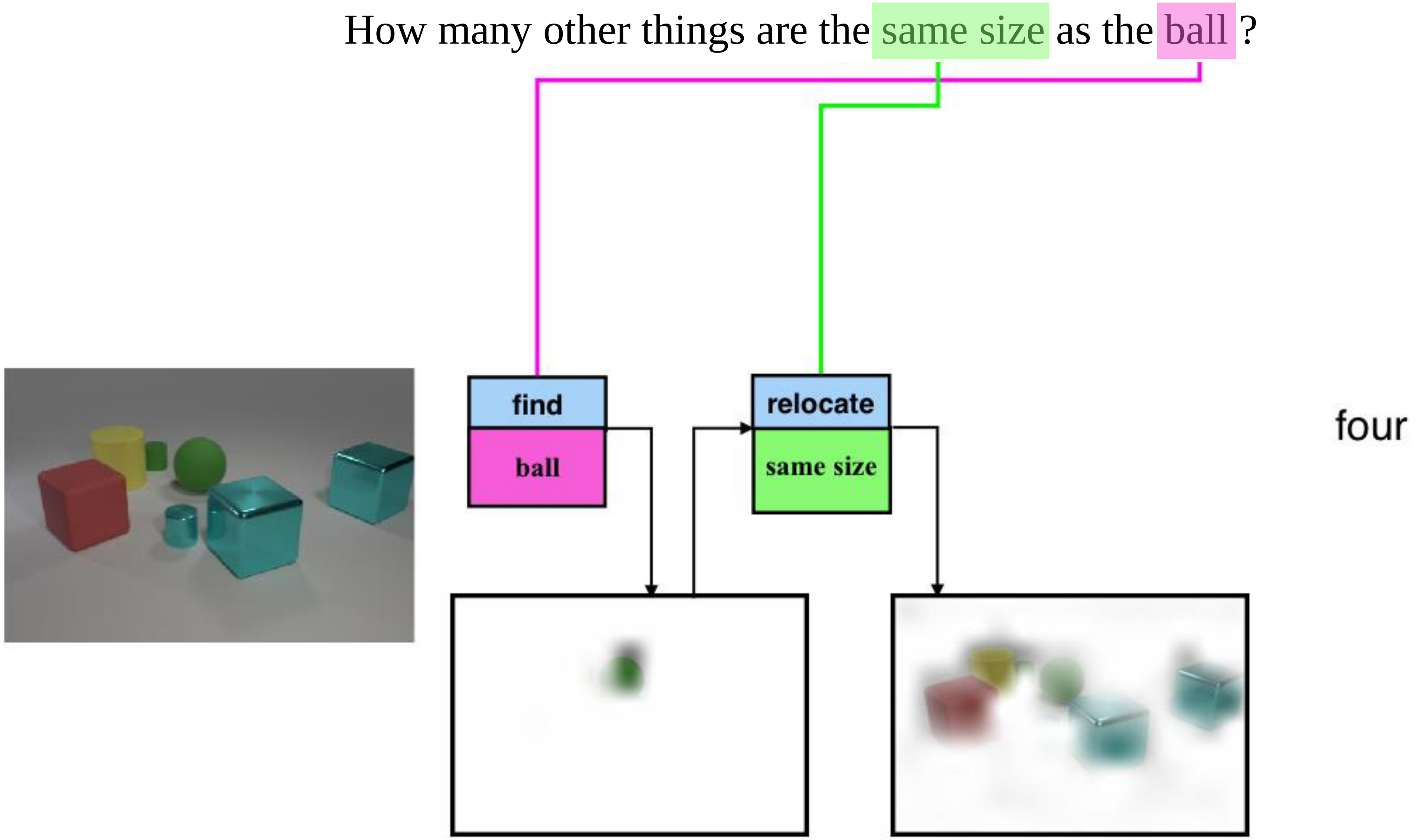
End-to-End Module Networks (N2NMN)



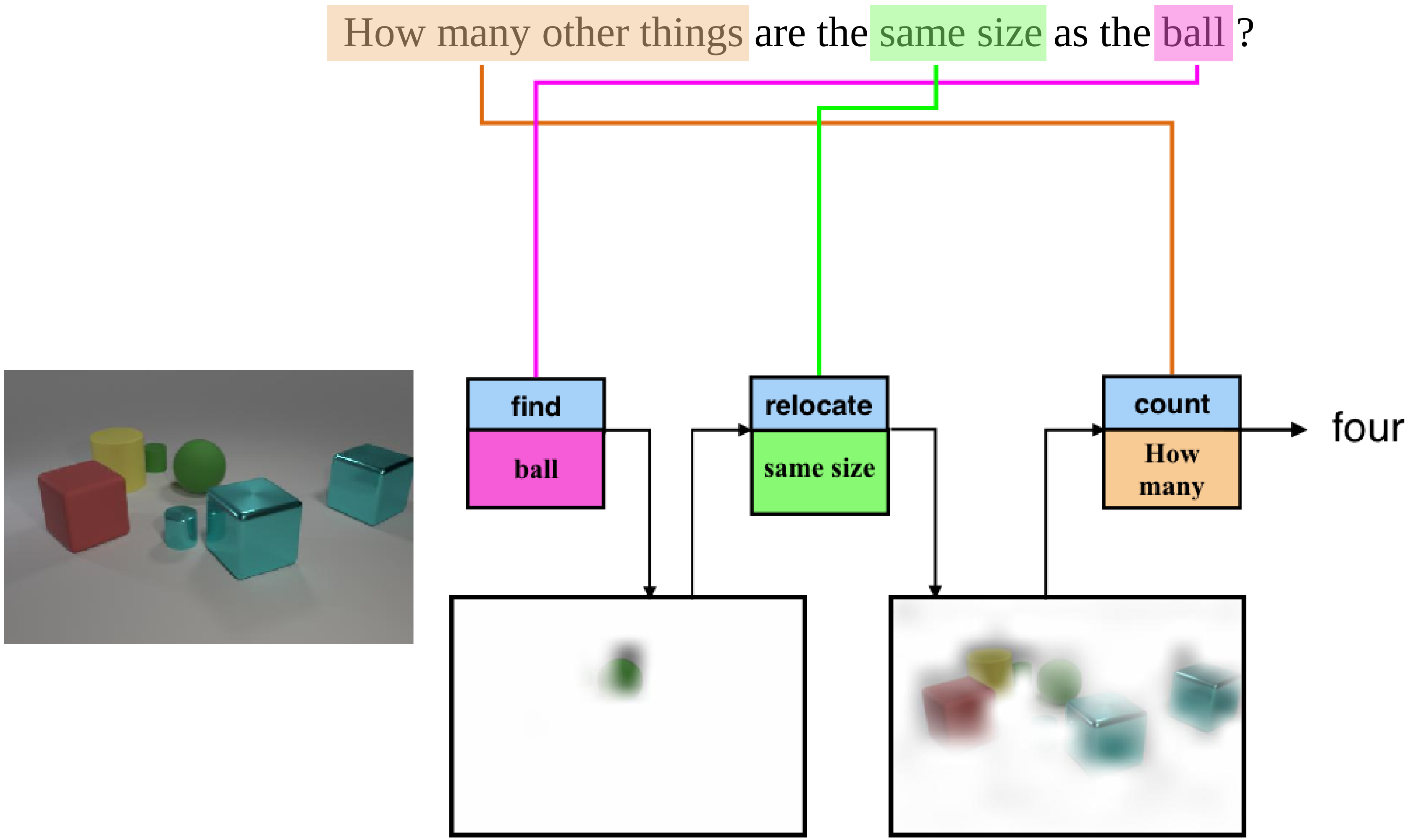
End-to-End Module Networks (N2NMN)



End-to-End Module Networks (N2NMN)

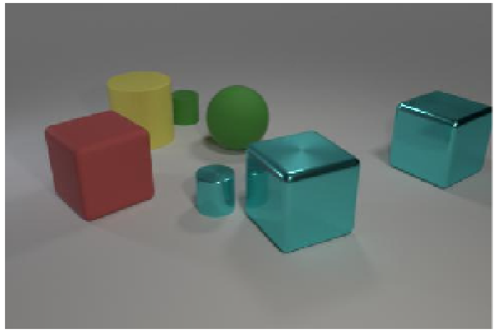


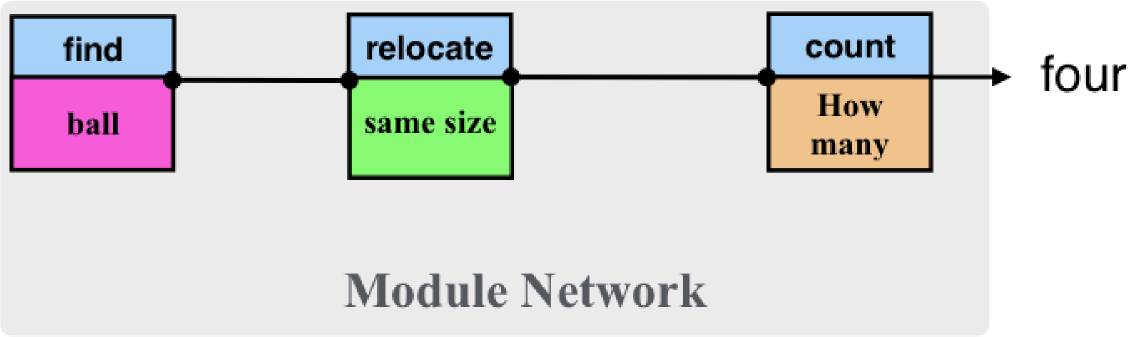
End-to-End Module Networks (N2NMN)



End-to-End Module Networks (N2NMN)

How many other things are the same size as the ball ?

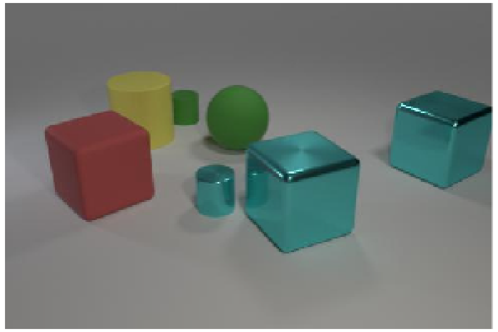


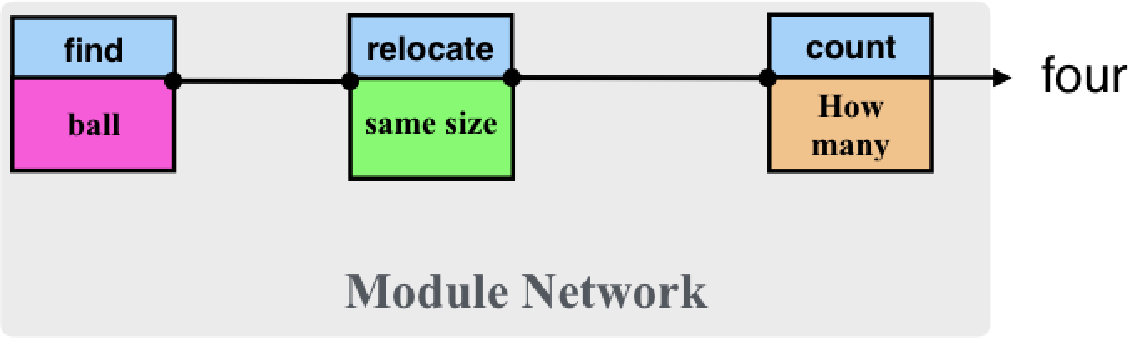


End-to-End Module Networks (N2NMN)

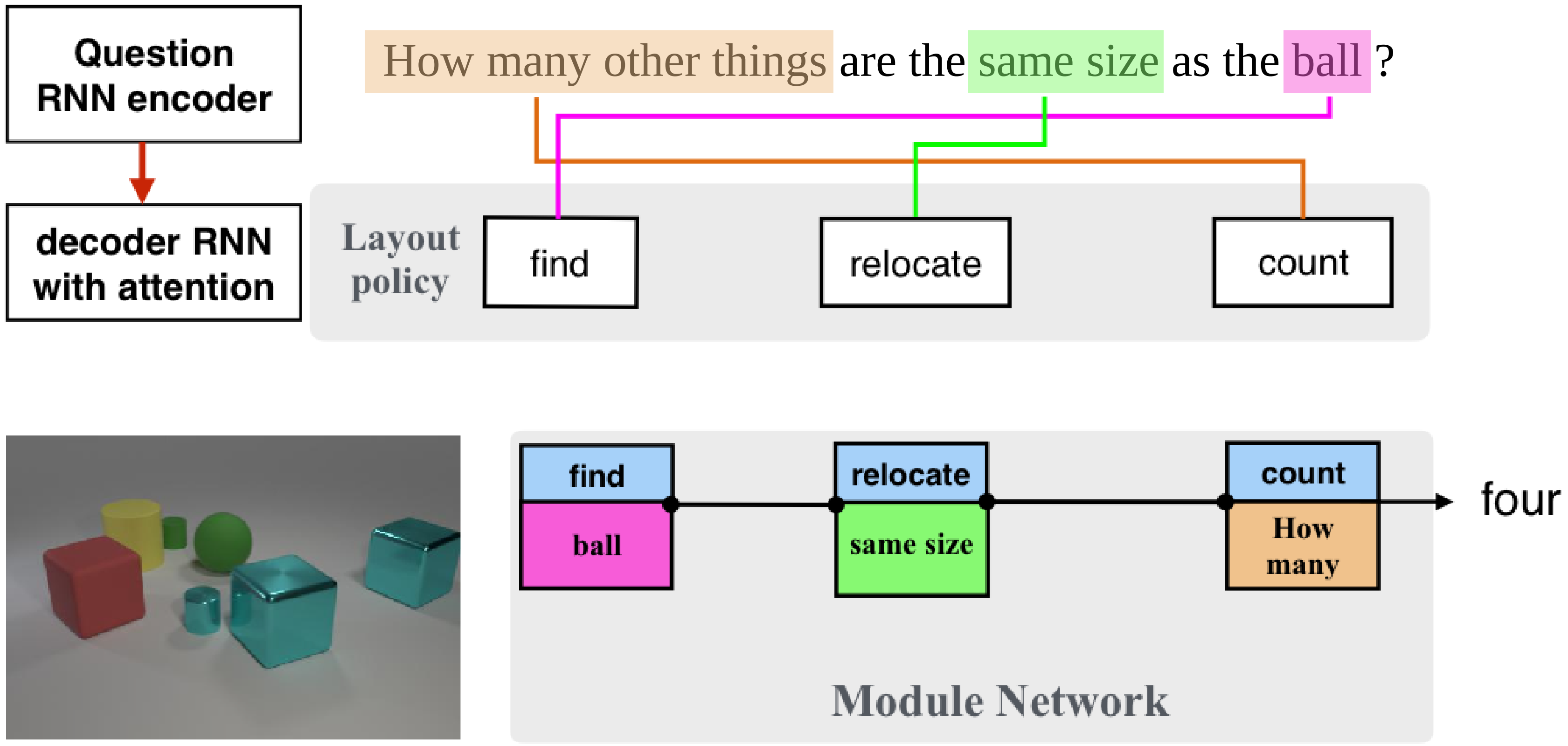
Question RNN encoder

How many other things are the same size as the ball ?

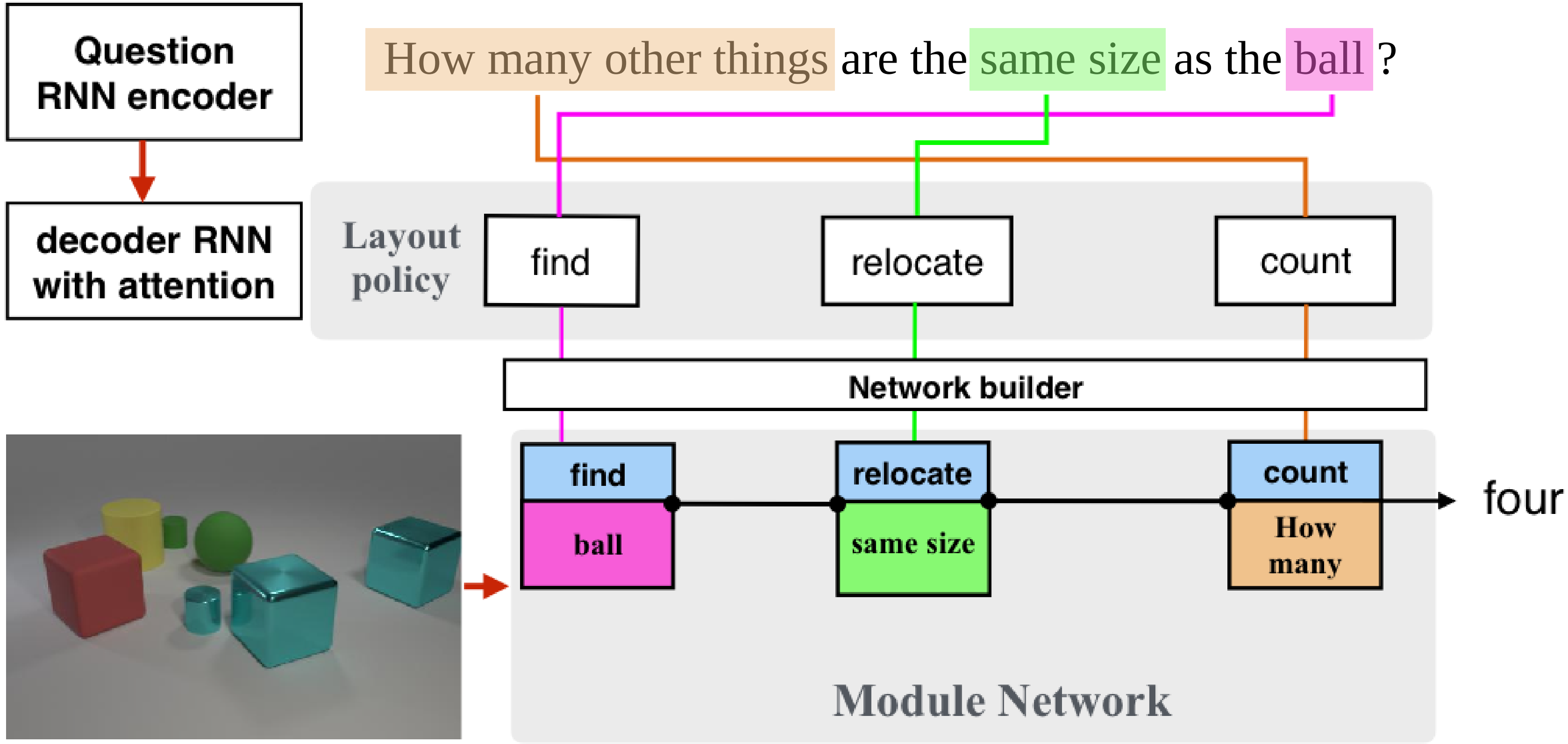




End-to-End Module Networks (N2NMN)



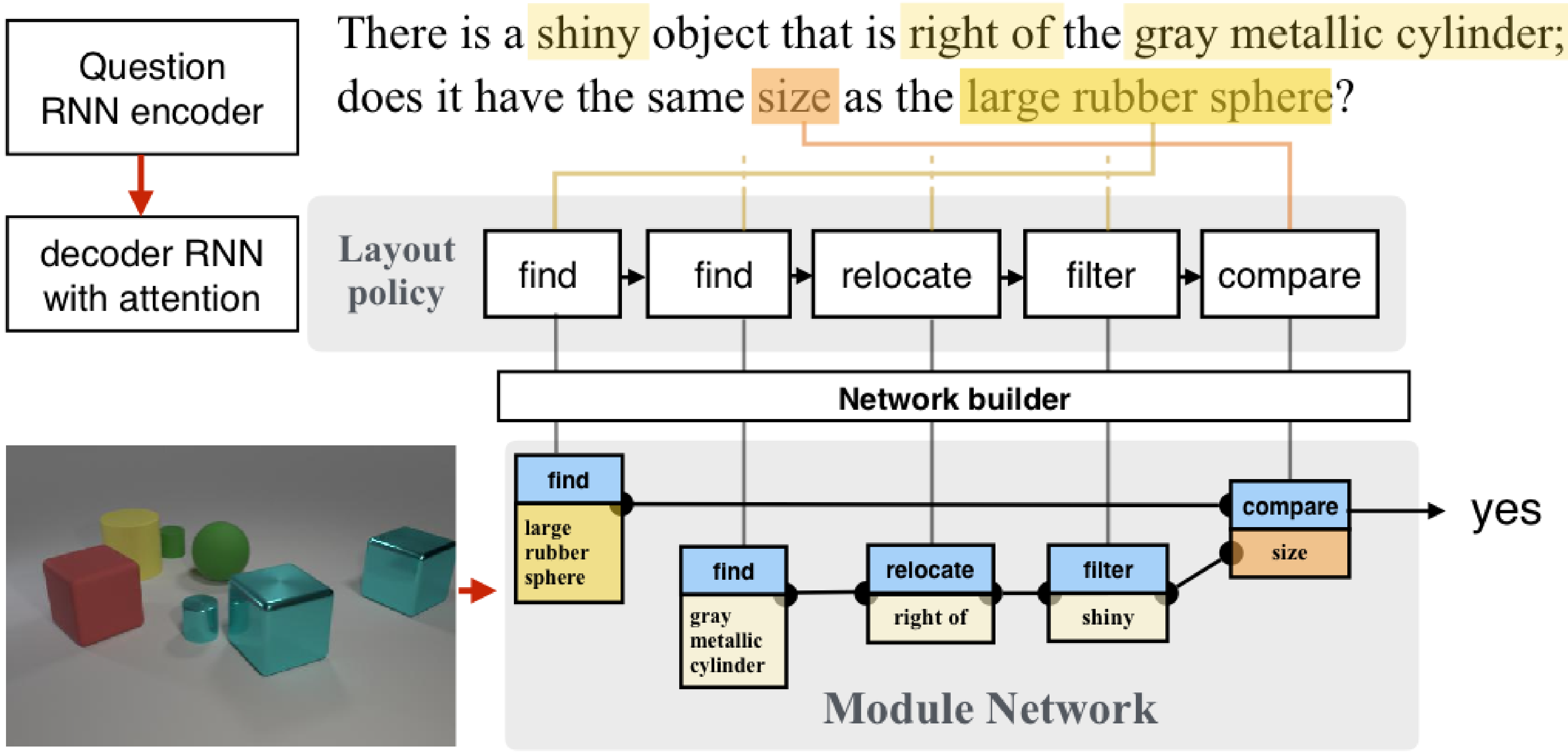
End-to-End Module Networks (N2NMN)



End-to-End Module Networks (N2NMN)

Layout policy:
what *tasks to do*

Neural modules:
how *to do them*



In this work, we simultaneously learn "what" and "how" end-to-end

Overview of N2NMN

Layout policy: ***what*** *tasks to do*

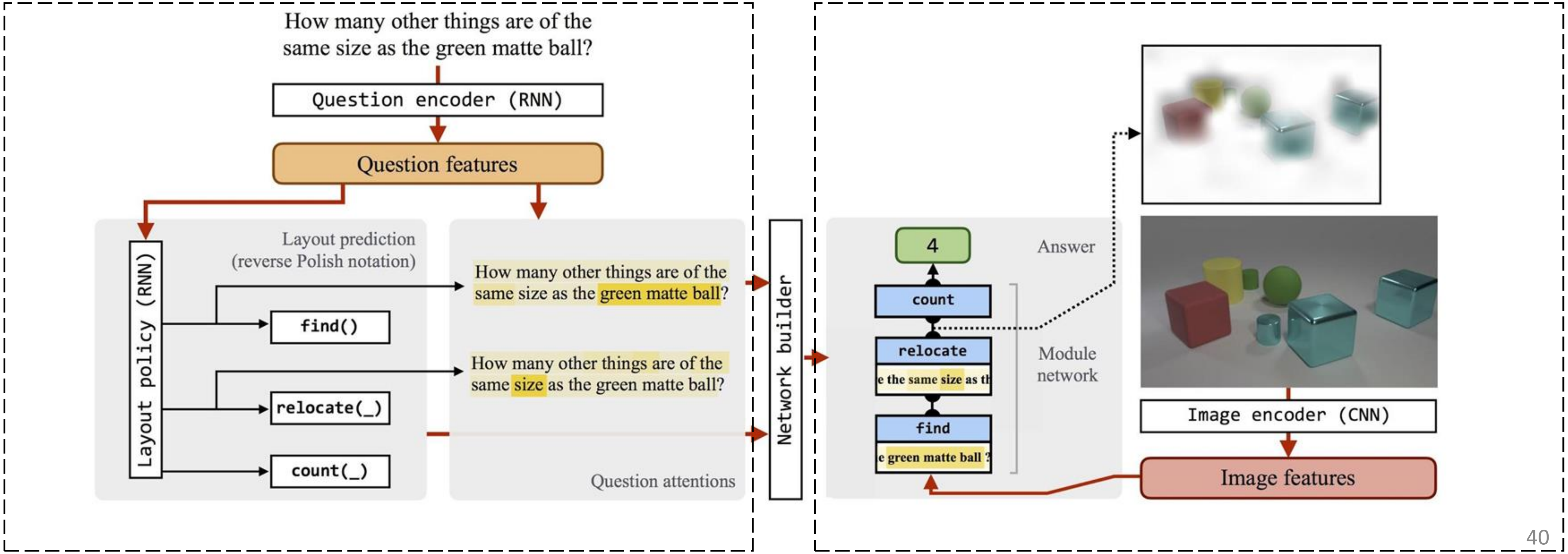
question -> *seq2seq RNN* -> module layout

`layout = [Find, Relocate, Count]`

Neural modules: ***how*** *to do them*

layout -> *dynamic networks* -> answer

`answer = "4"`



Layout Policy: Question -> Dynamic Networks

Question

"How many other things are of the same size as the green matte ball?"

Translate questions to layout tokens (similar to machine translation)

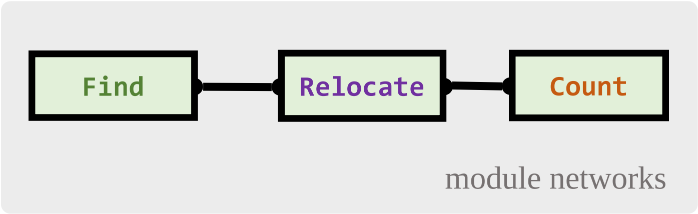
Layout tokens

```
layout = [find, relocate, count]
```

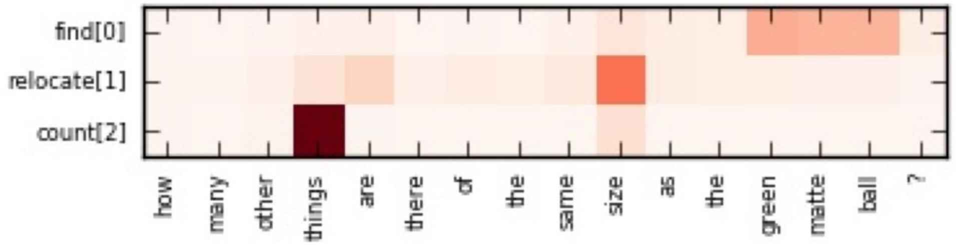
assemble dynamic networks

Module layout

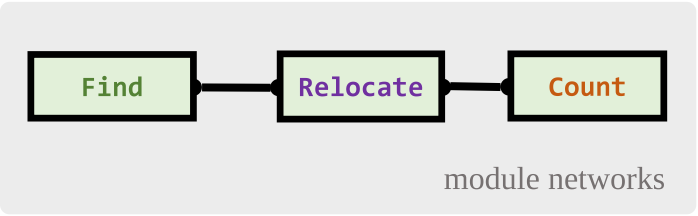
```
count(relocate(find()))
```



Textual Attention in seq2seq RNNs



`count(relocate(find()))`



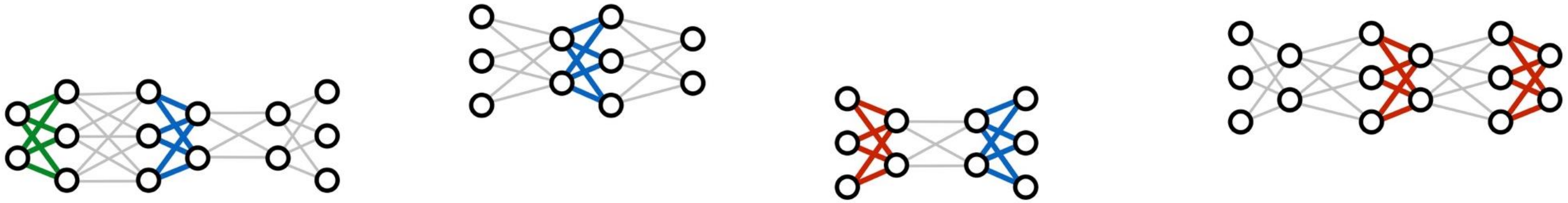
"How many other things are of the same size as the green matte ball?"

Module Networks

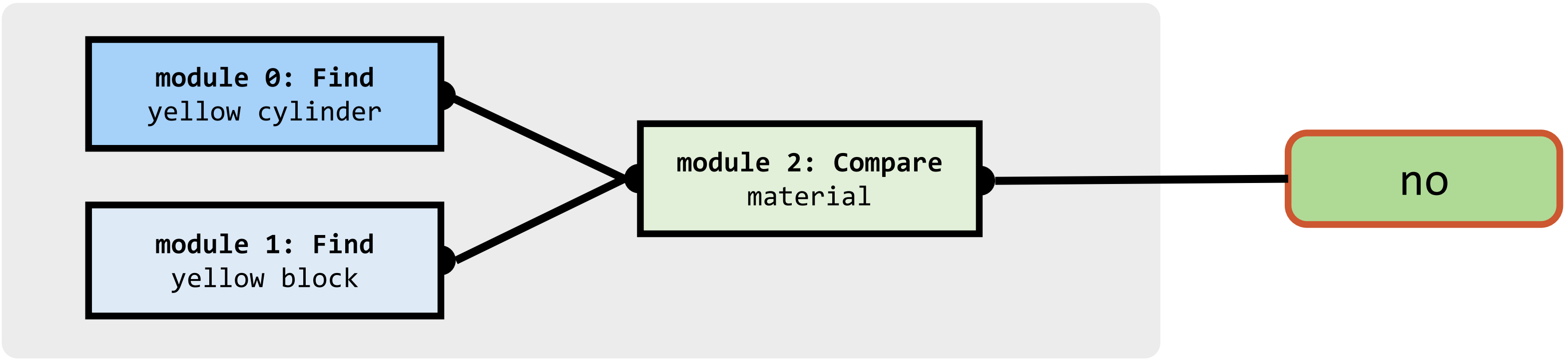
- Modules can be added as needed for a given problem

Module name	Att-inputs	Features	Output	Implementation details
`find`	(none)	x_{vis}, x_{txt}	att	$a_{out} = \text{conv}_2\left(\text{conv}_1(x_{vis}) \odot W x_{txt}\right)$
`relocate`	a	x_{vis}, x_{txt}	att	$a_{out} = \text{conv}_2\left(\text{conv}_1(x_{vis}) \odot W_1 \text{sum}(a \odot x_{vis}) \odot W_2 x_{txt}\right)$
`and`	a_1, a_2	(none)	att	$a_{out} = \text{minimum}(a_1, a_2)$
`or`	a_1, a_2	(none)	att	$a_{out} = \text{maximum}(a_1, a_2)$
`filter`	a	x_{vis}, x_{txt}	att	$a_{out} = \texttt{and}(a, \texttt{find}[x_{vis}, x_{txt}]())$, *i.e.* reusing `find` and `and`
`[exist, count]`	a	(none)	ans	$y = W^T \text{vec}(a)$
`describe`	a	x_{vis}, x_{txt}	ans	$y = W_1^T \left(W_2 \text{sum}(a \odot x_{vis}) \odot W_3 x_{txt}\right)$
`[eq_count, more, less]`	a_1, a_2	(none)	ans	$y = W_1^T \text{vec}(a_1) + W_2^T \text{vec}(a_2)$
`compare`	a_1, a_2	x_{vis}, x_{txt}	ans	$y = W_1^T \left(W_2 \text{sum}(a_1 \odot x_{vis}) \odot W_3 \text{sum}(a_2 \odot x_{vis}) \odot W_4 x_{txt}\right)$

- Modules are dynamically assembled into networks on-the-fly



Learning from Expert Layouts



Stage 1: train the model to predict the ground-truth (gold) layout with supervised learning (behavioral cloning from expert layouts)

is the material of the yellow block same as the yellow cylinder?

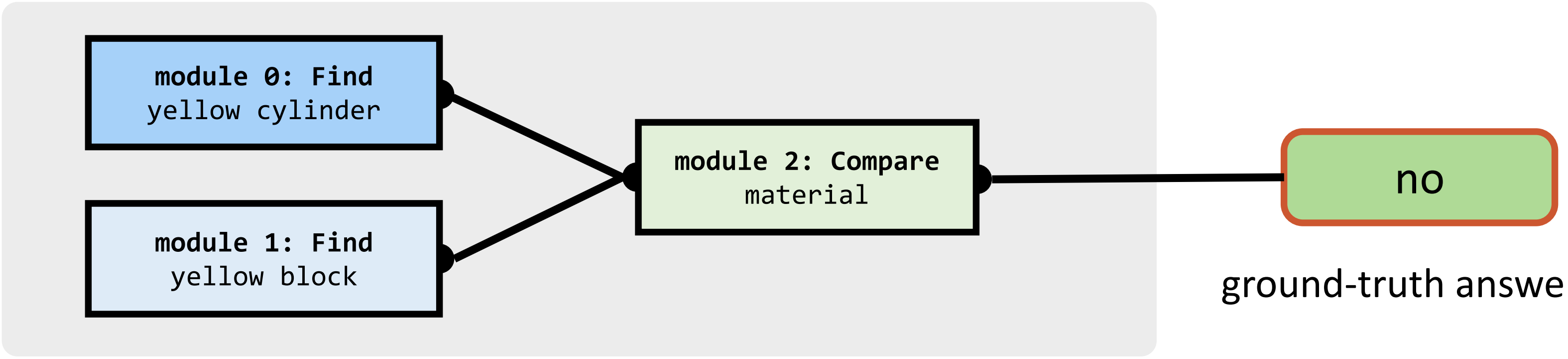
layout policy
(seq2seq RNN)

`Expert (Gold) Layout: [Find, Find, Compare]`

Expert (gold) layout from
- dataset annotations *or*
- syntactic parsing

End-to-End Layout Search with Policy Gradients

is the material of the yellow block the same as the yellow cylinder?

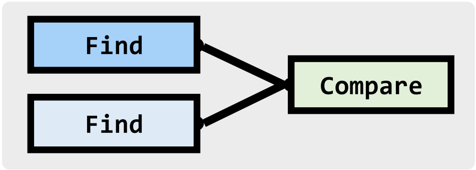


ground-truth answer: “no”

Stage 2: sample multiple candidate layouts from the layout policy, and optimize with *policy gradient* (REINFORCE)

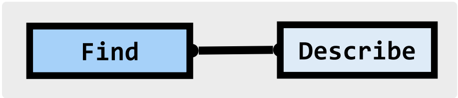
layout policy

```
Sampled Layout 1: [Find, Find, Compare]
```



prediction: “no”
QA loss = 0.05

```
Sampled Layout 2: [Find, Describe]
```



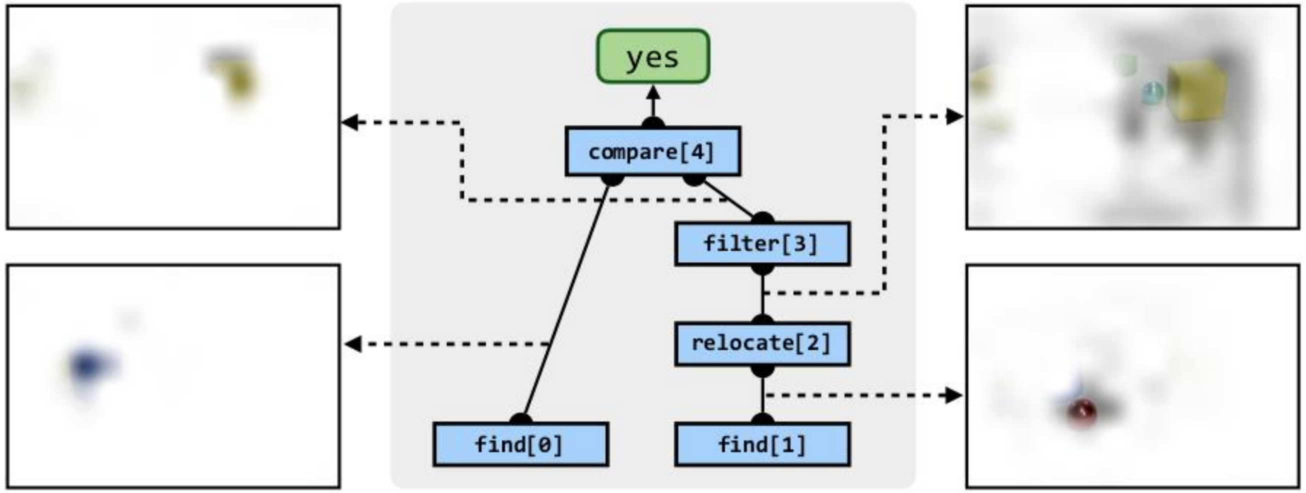
prediction: “yes”
QA loss = 1.72

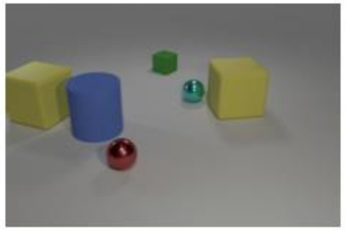
```
Sampled Layout 3: [Find, Transform, Describe]
```

prediction: “no”
QA loss = 0.08

...

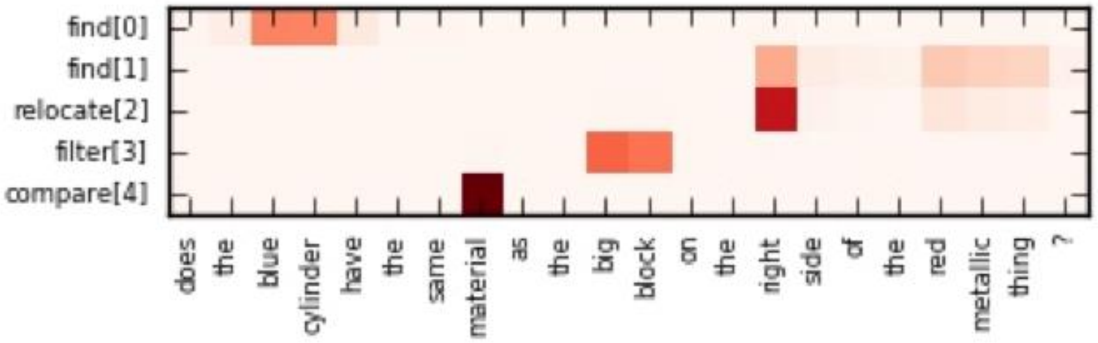
Qualitative results on the CLEVR dataset (synthetic images)





Does the blue cylinder have the same material as the big block on the right side of the red metallic thing?

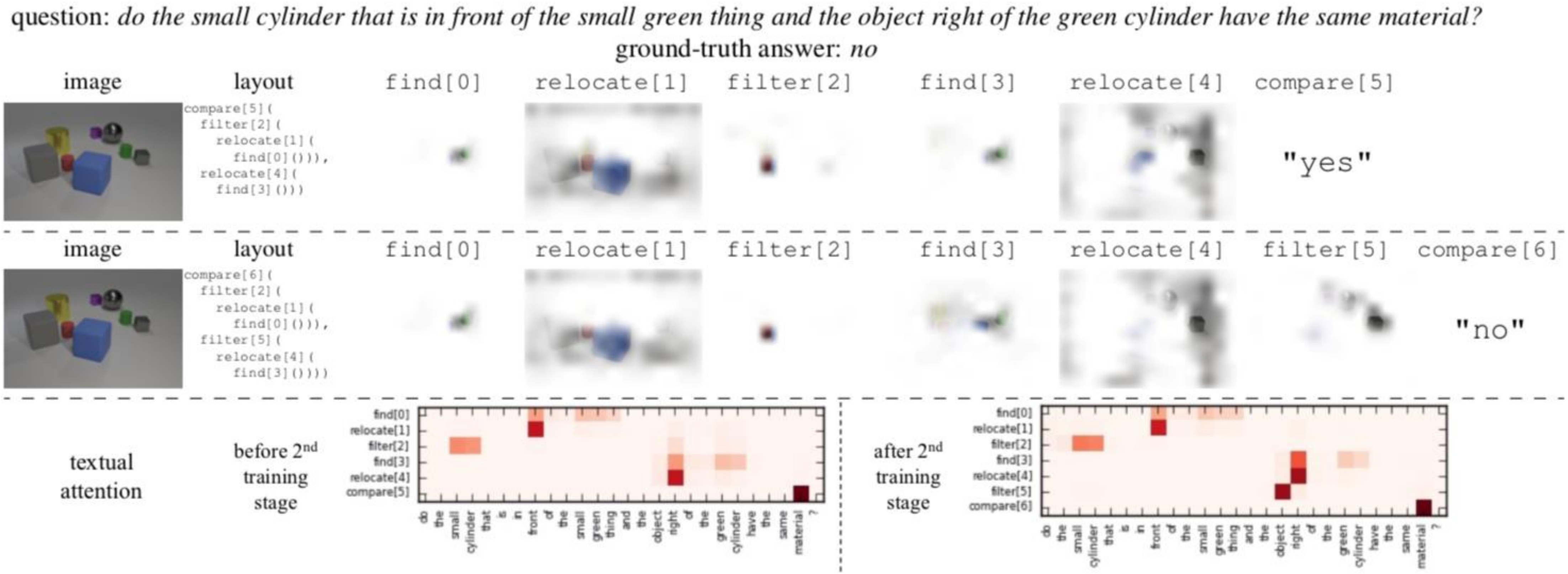
textual attention for each module



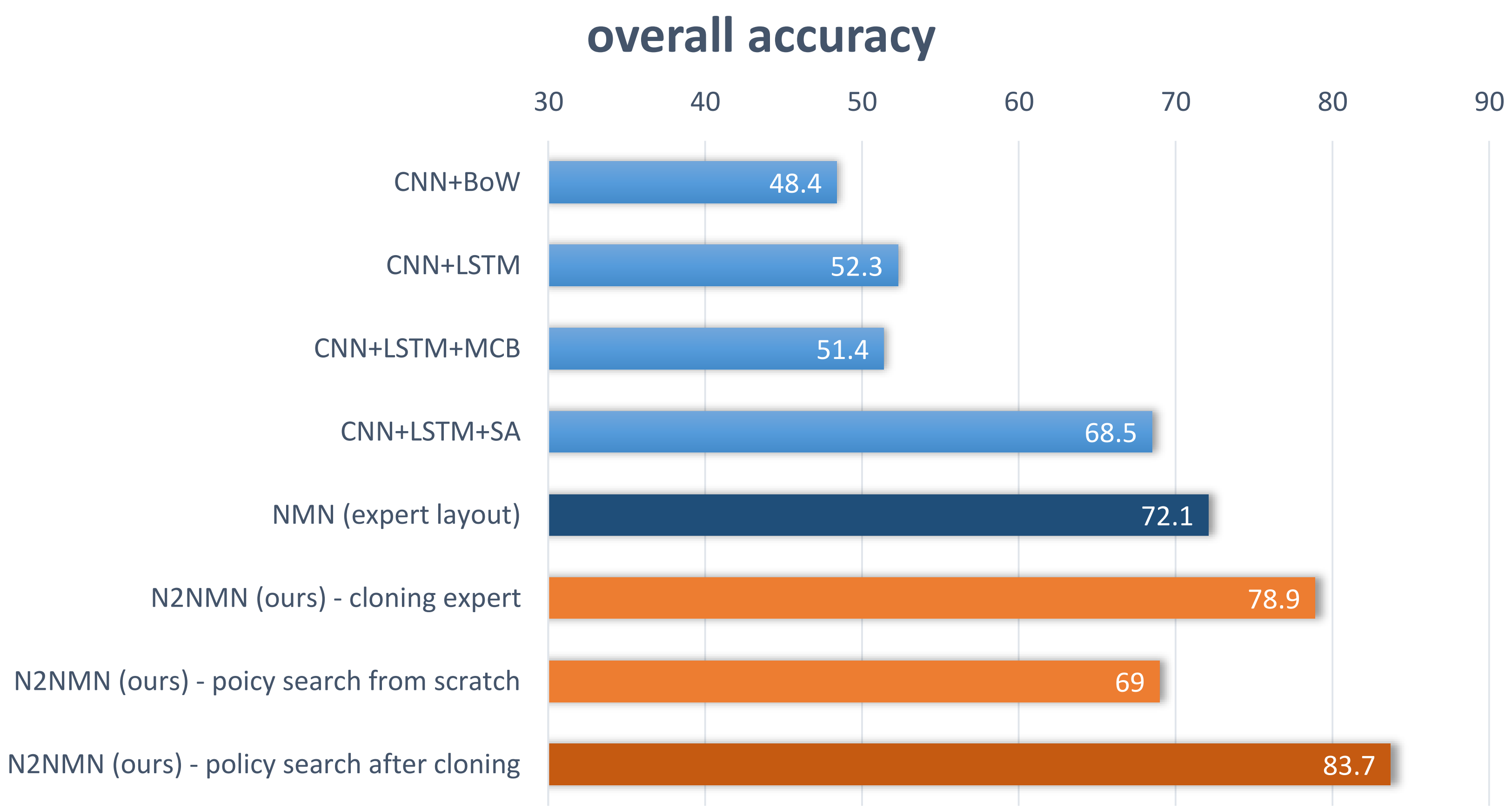
Qualitative results on the CLEVR dataset (synthetic images)

Stage 1
clone expert
(gold) layout

Stage 2
end-to-end
layout search

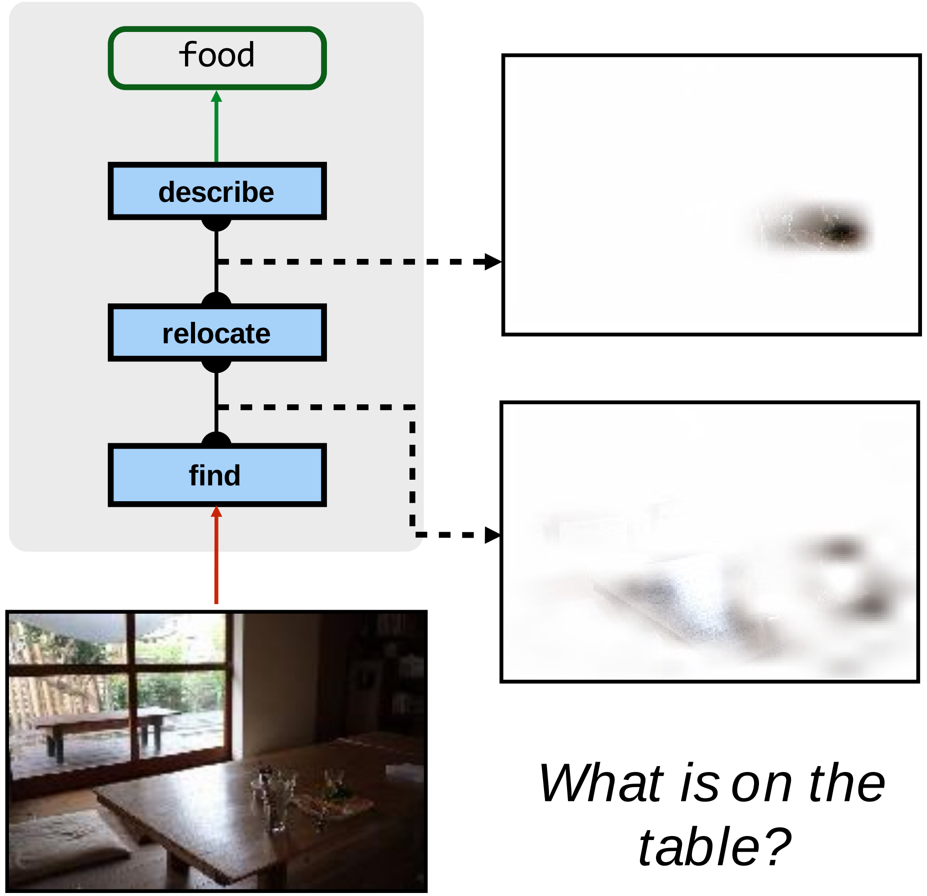


Quantitative results on the CLEVR dataset (synthetic images)



- Superior performance with end-to-end training

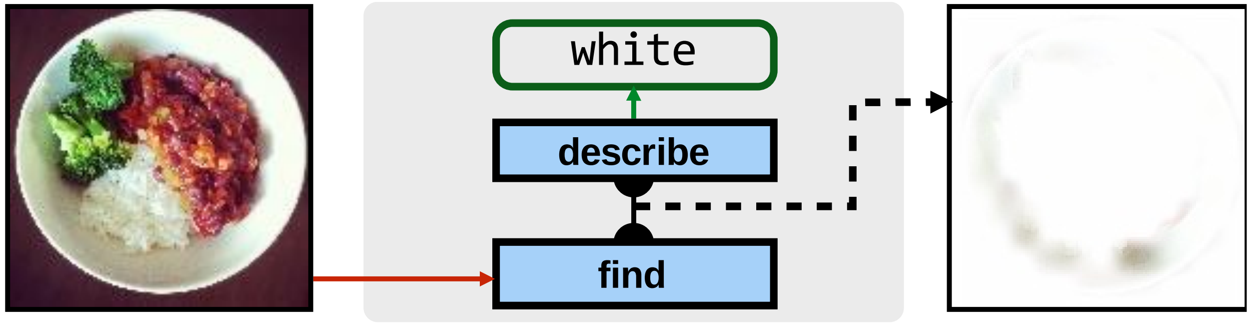
Qualitative results on the VQA dataset (natural images)



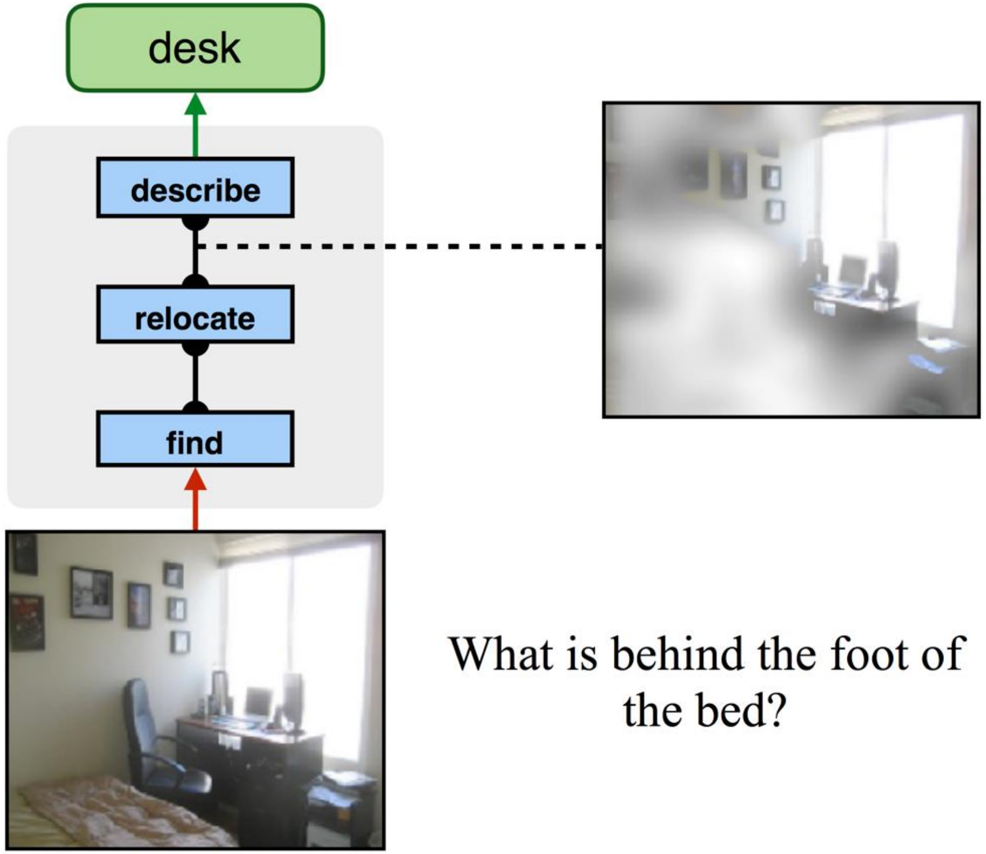
What is on the table?

Qualitative results on the VQA dataset (natural images)

What color is the plate?

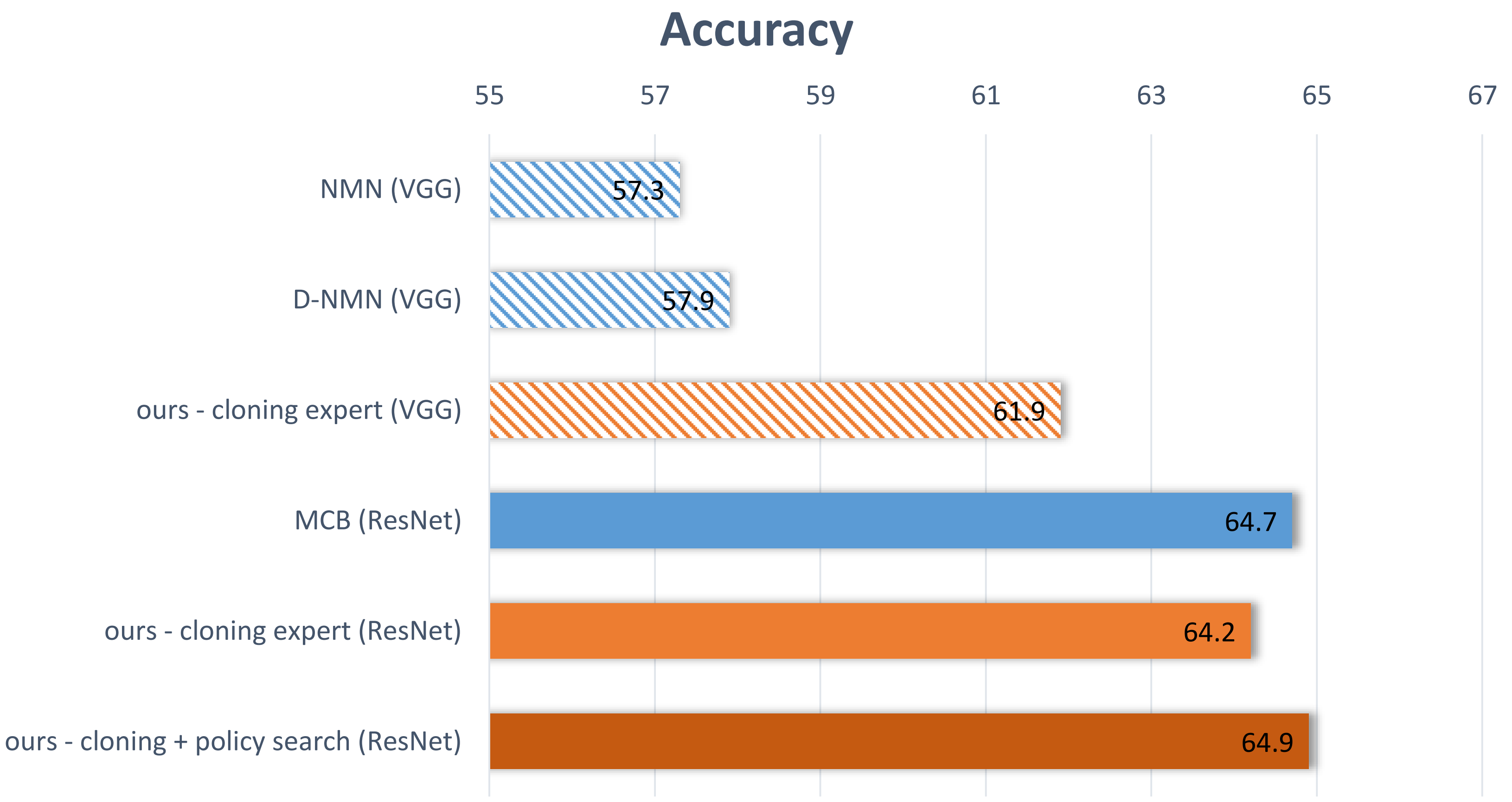


Qualitative results on the VQA dataset (natural images)



What is behind the foot of the bed?

Quantitative results on the VQA dataset (natural images)

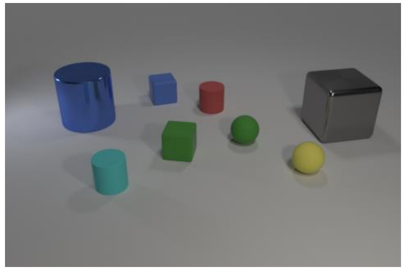


- Works well on real images and questions

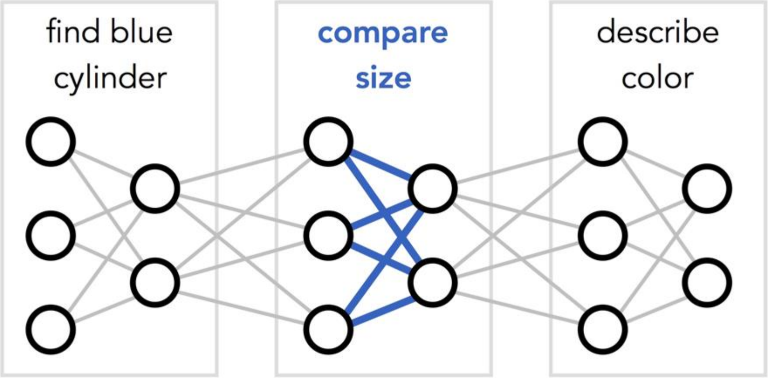
Summary of N2NMN

- Discrete compositionality with trainable and reusable modules
- Jointly train policy (*what*) and compositional modules (*how*)
- A possible way to bridge neural + symbolic

What color is the thing with the same size as the blue cylinder?



```
def answer_this_question(image):
    object_1 = find(image, 'blue cylinder')
    object_2 = compare(object_1, 'size')
    answer = describe(object_2, 'color')
    return answer
```



output answer: “gray”

Fine-grained Textual Explanations

Trevor Darrell
UC Berkeley

With **Lisa Anne Hendricks**, Zeynep Akata, Ronghang Hu, Bernt Schiele, Marcus Rohrbach, …

Cardinal


What type of bird is this?
It is a **Cardinal** because it is a **red bird** with a **red beak** and a **black face**

Explanations: Generating Natural Language Explanations of Visual Decisions

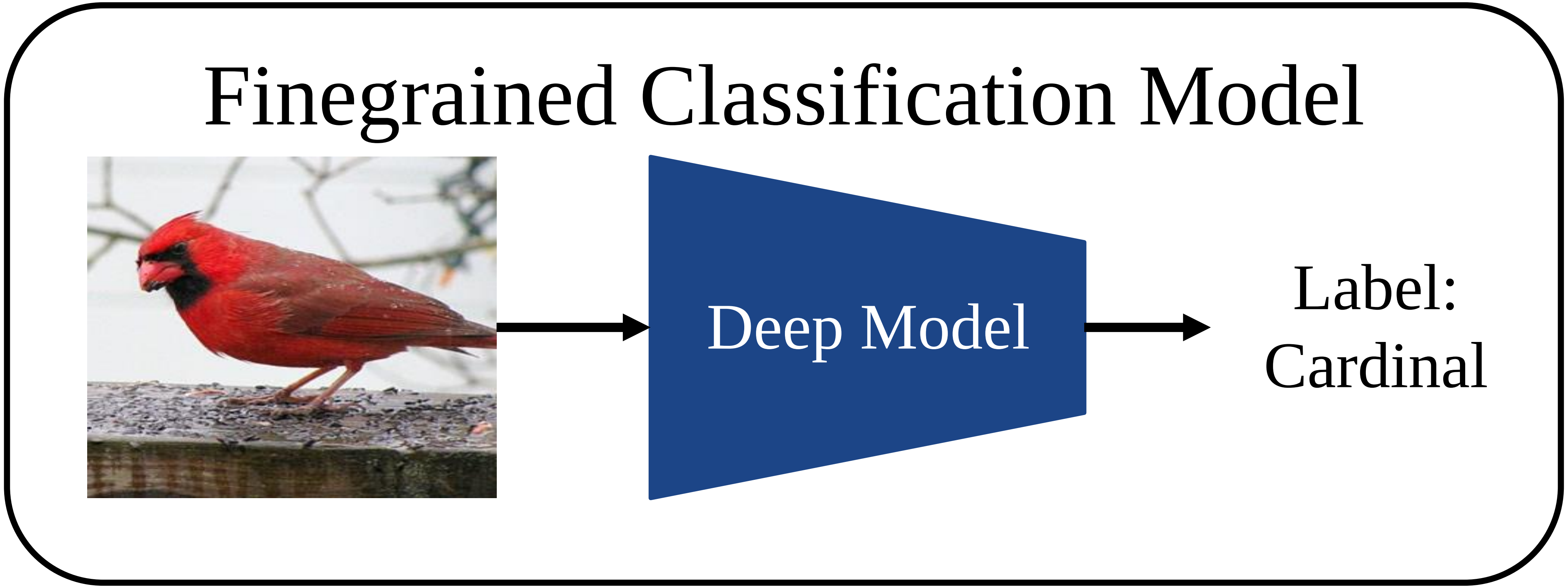
Explanations

This is a ***White Necked Raven*** because it is a black bird with a white nape and a large beak.

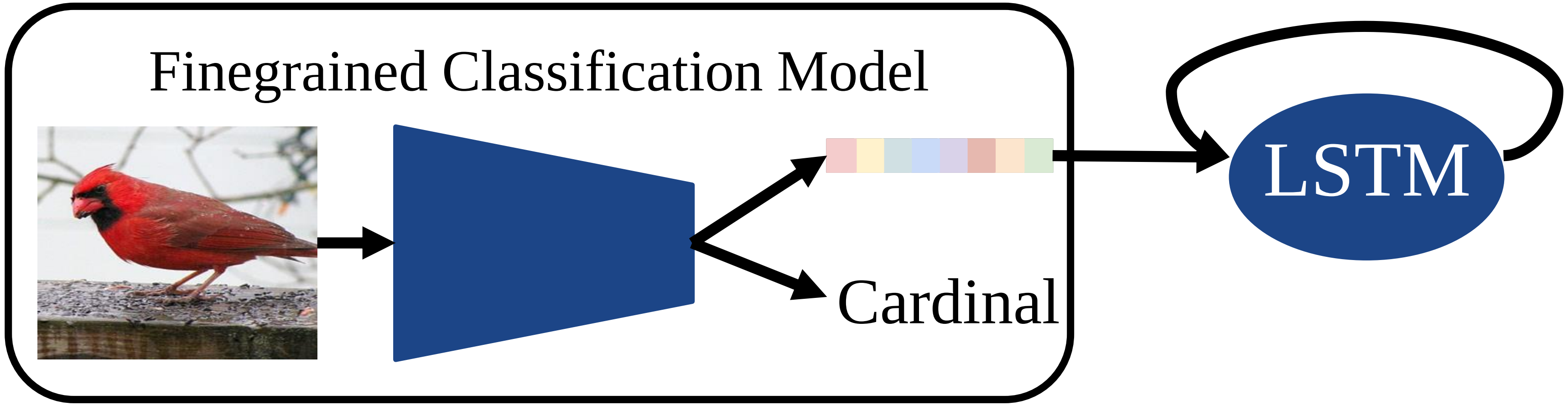
Hendricks et al. Generating Visual Explanations. ECCV 2016.
Park, **Hendricks** et al. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. CVPR 2017.
Hendricks et al. Generating Counterfactual Explanations with Natural Language. ICML Workshops 2018.
Hendricks et al. Grounding Visual Explanations. ECCV 2018.

Fine-grained Explanations

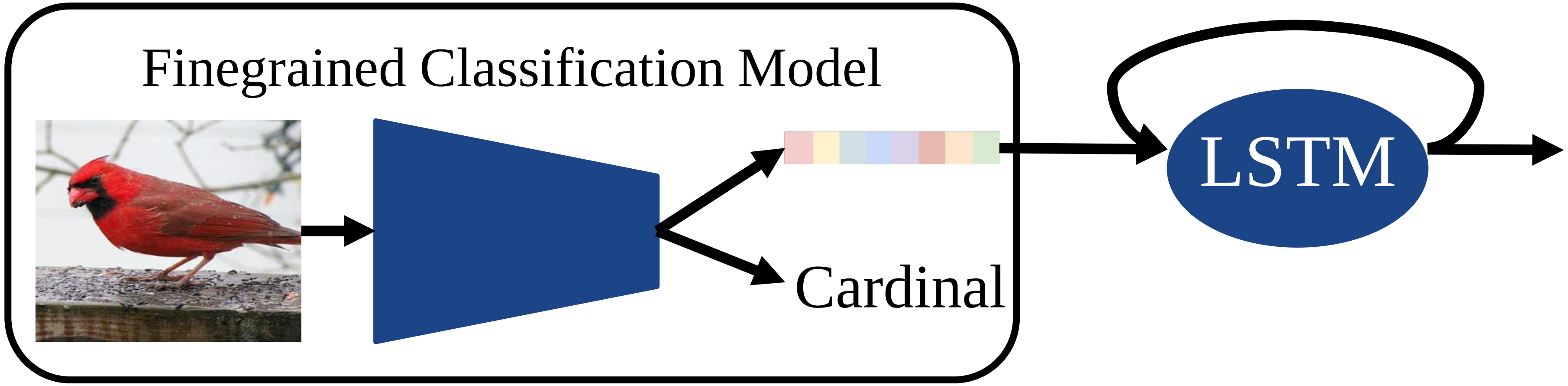
Fine-grained Explanations



Fine-grained Explanations



Fine-grained Explanations



This is a ***Cardinal*** because this bird has a red crown, a short bill, and a red breast.

Descriptions from: Reed et al. Learning deep representations of finegrained visual descriptions. CVPR 2016.

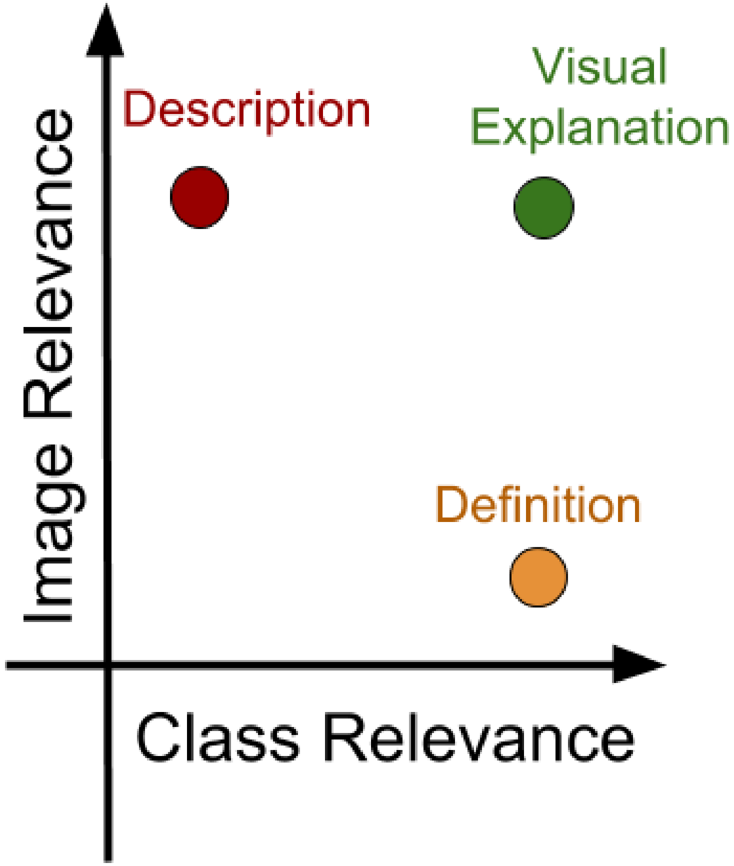
Fine-grained Explanations

This is a ***Cardinal*** because this bird has a red crown, a short bill, and a red breast.

Fine-grained Explanations

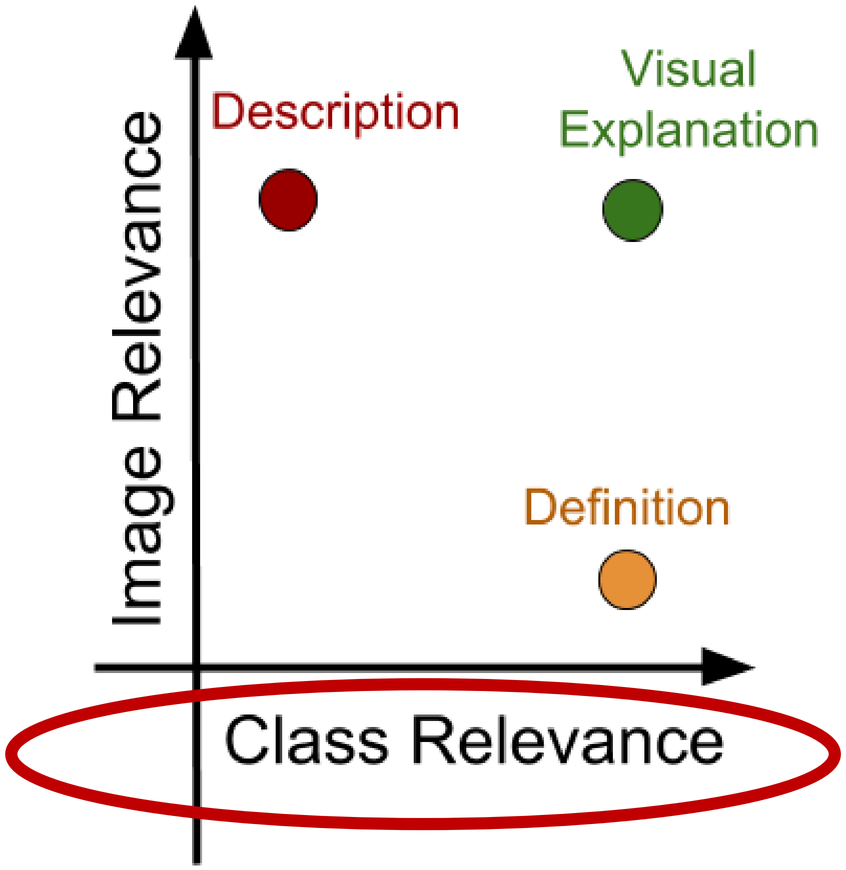
This is a ***Cardinal*** because this bird has a red crown, a short bill, and a red breast.

What makes a good visual explanation?



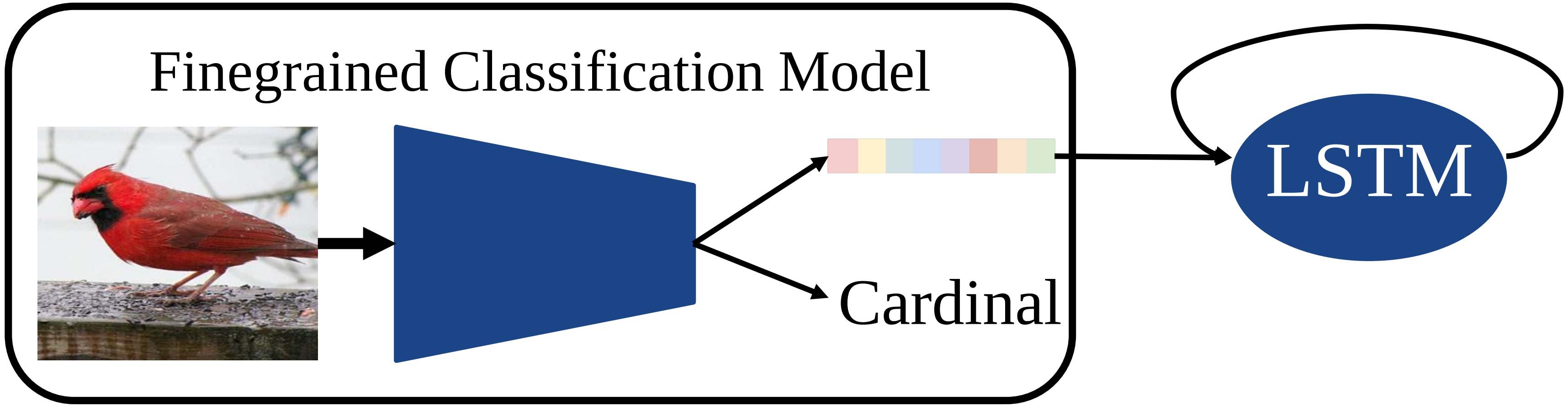
This is a ***Cardinal*** because this bird has a red crown, a short bill, and a red breast.

What makes a good visual explanation?

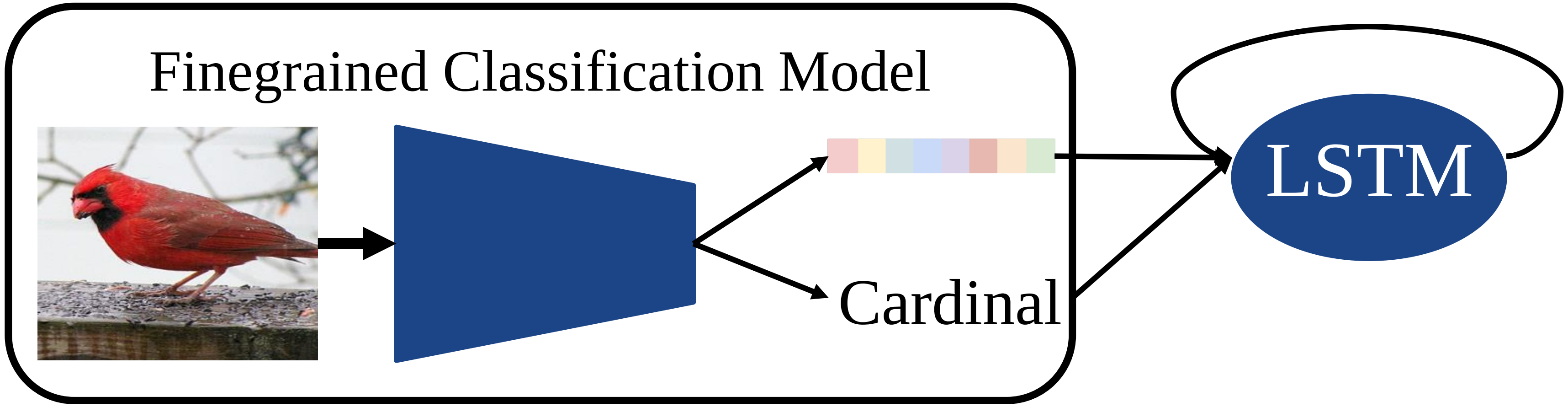


This is a ***Cardinal*** because this is a red bird with a black face and a red beak.

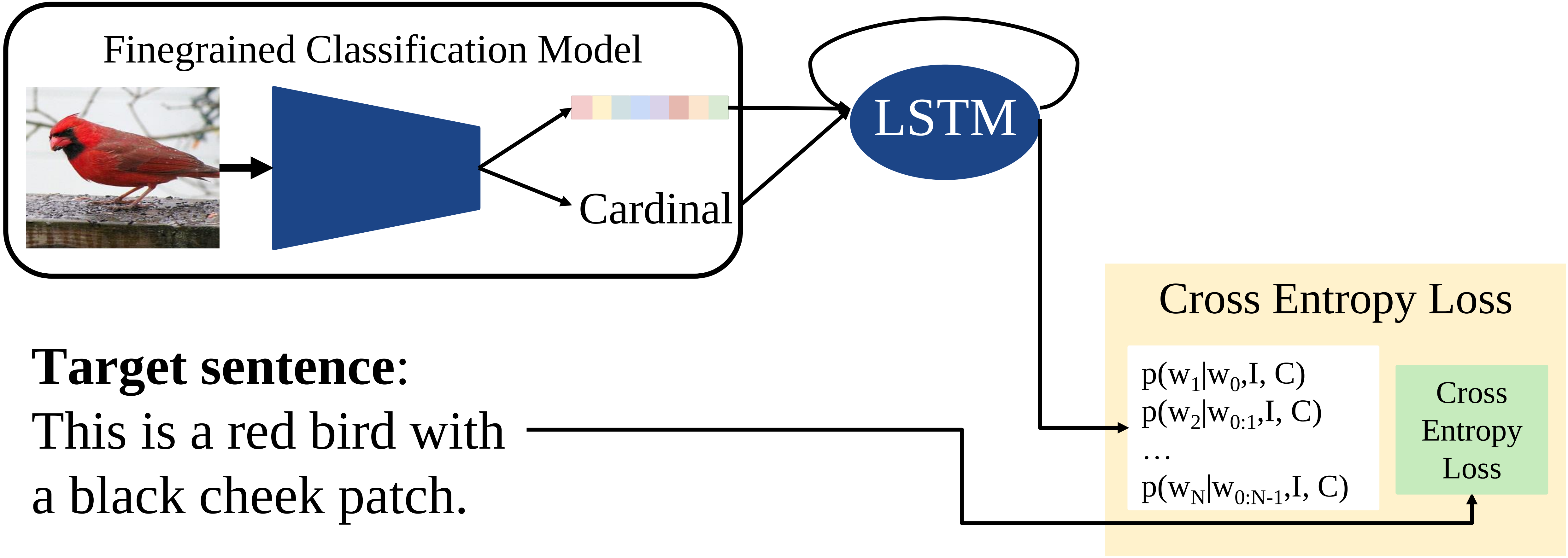
Visual Explanation Model



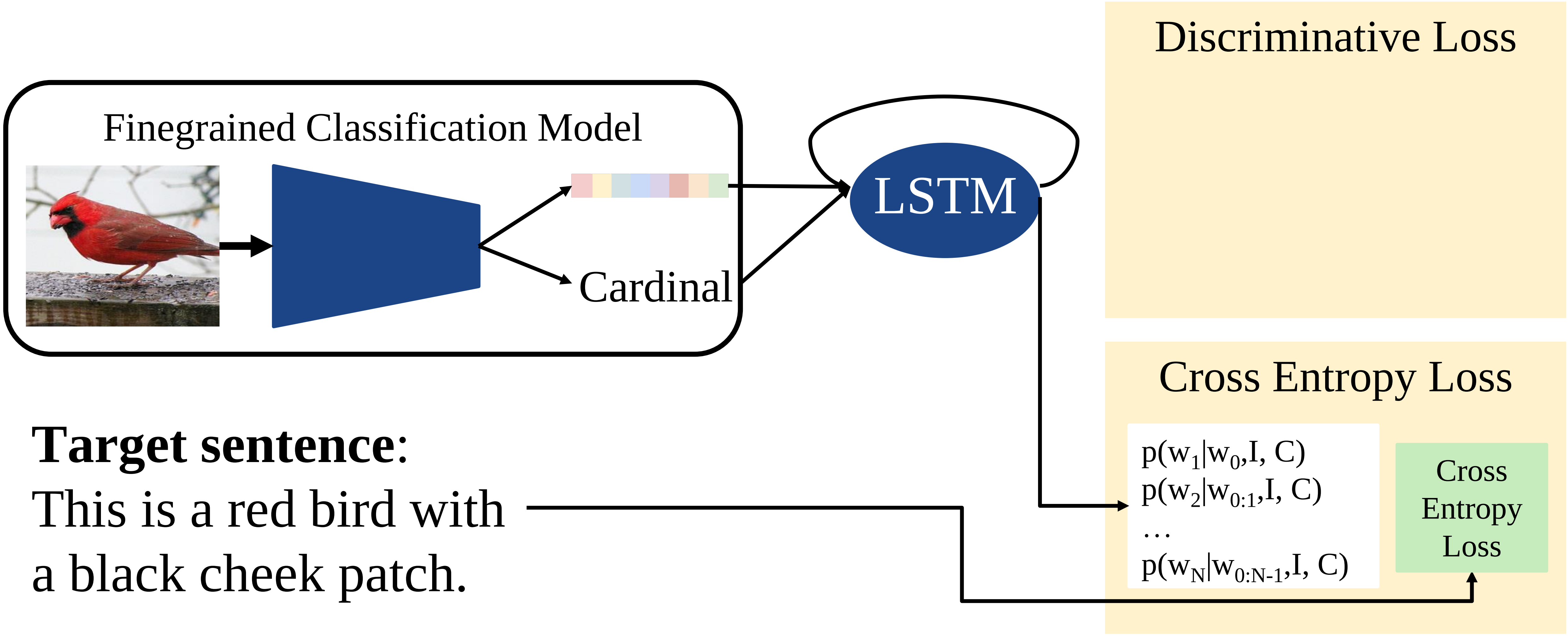
Visual Explanation Model



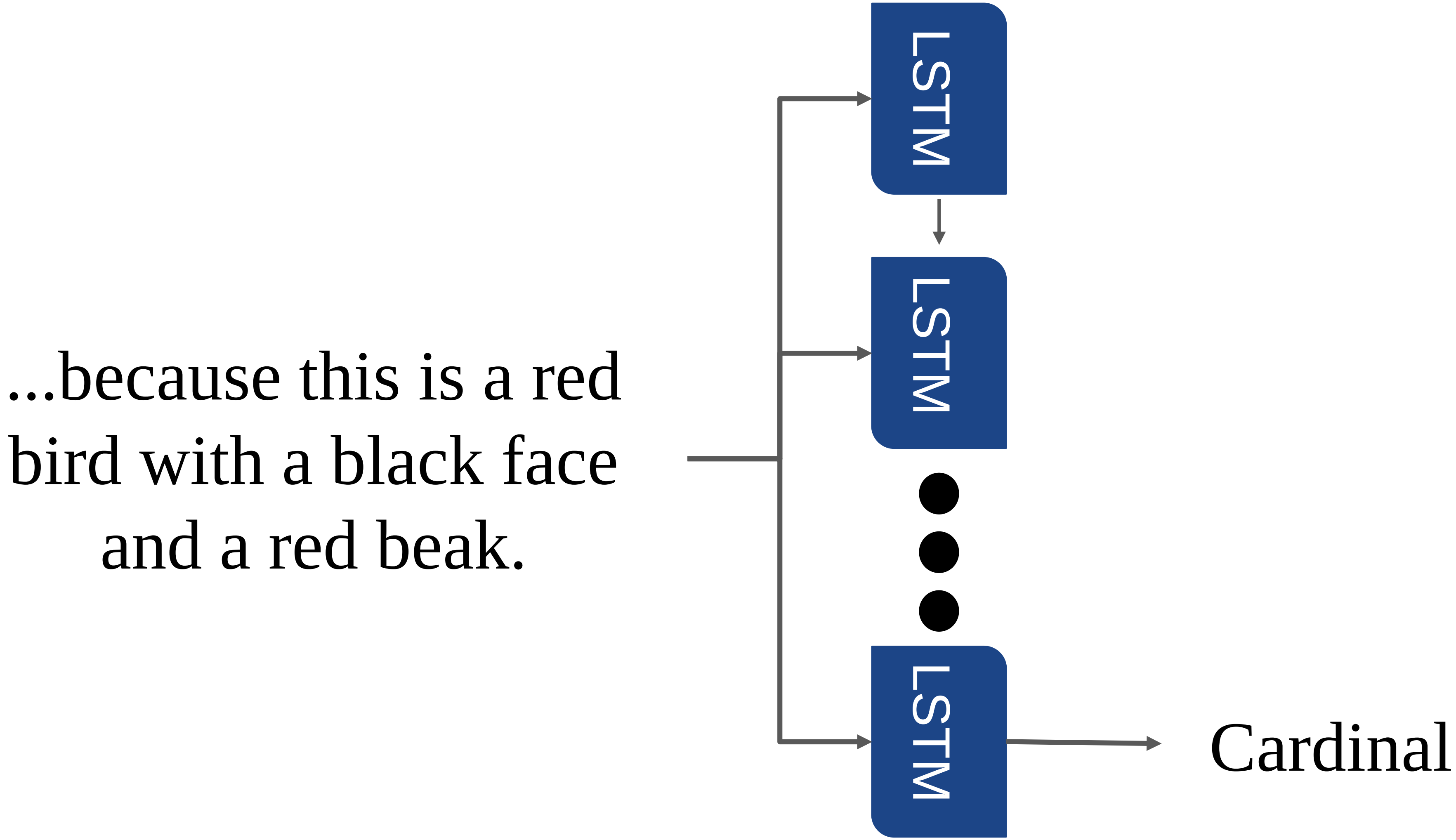
Visual Explanation Model



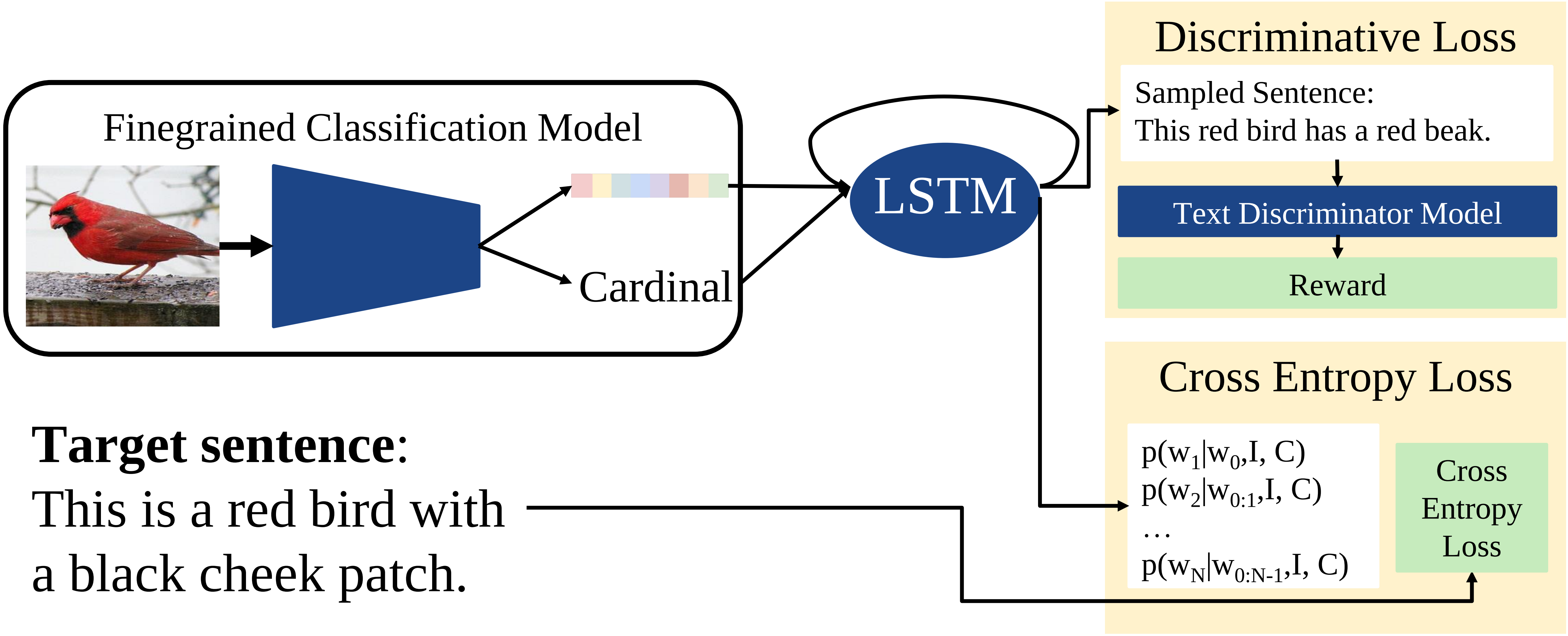
Visual Explanation Model



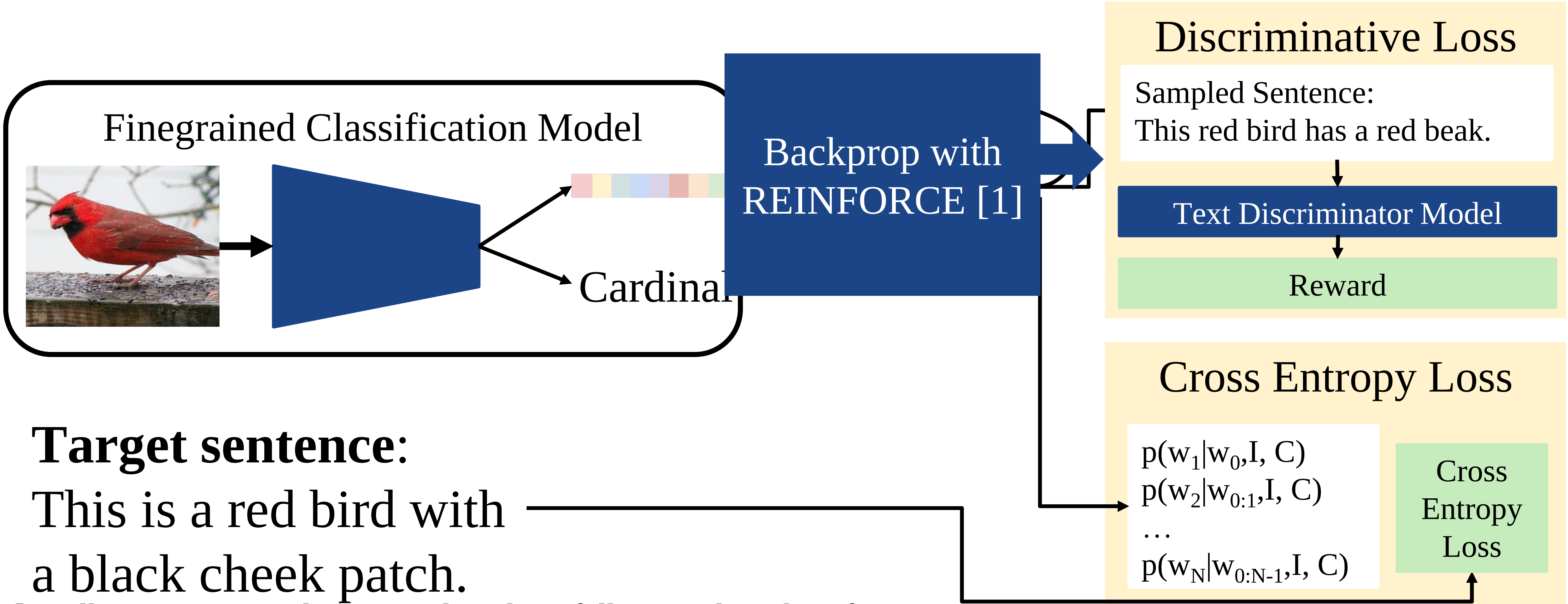
Text Discriminator Model



Visual Explanation Model



Visual Explanation Model



[1] Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning (1992)

This is a ***White Necked Raven...***

Description: because this bird is nearly all black with a short pointy bill.

This is a ***White Necked Raven...***

Description: because this bird is nearly all black with a short pointy bill.

Explanation: because this is a black bird with a white nape and a large black beak.

Evaluating Explanations

Choose the image which most closely matches the following text:

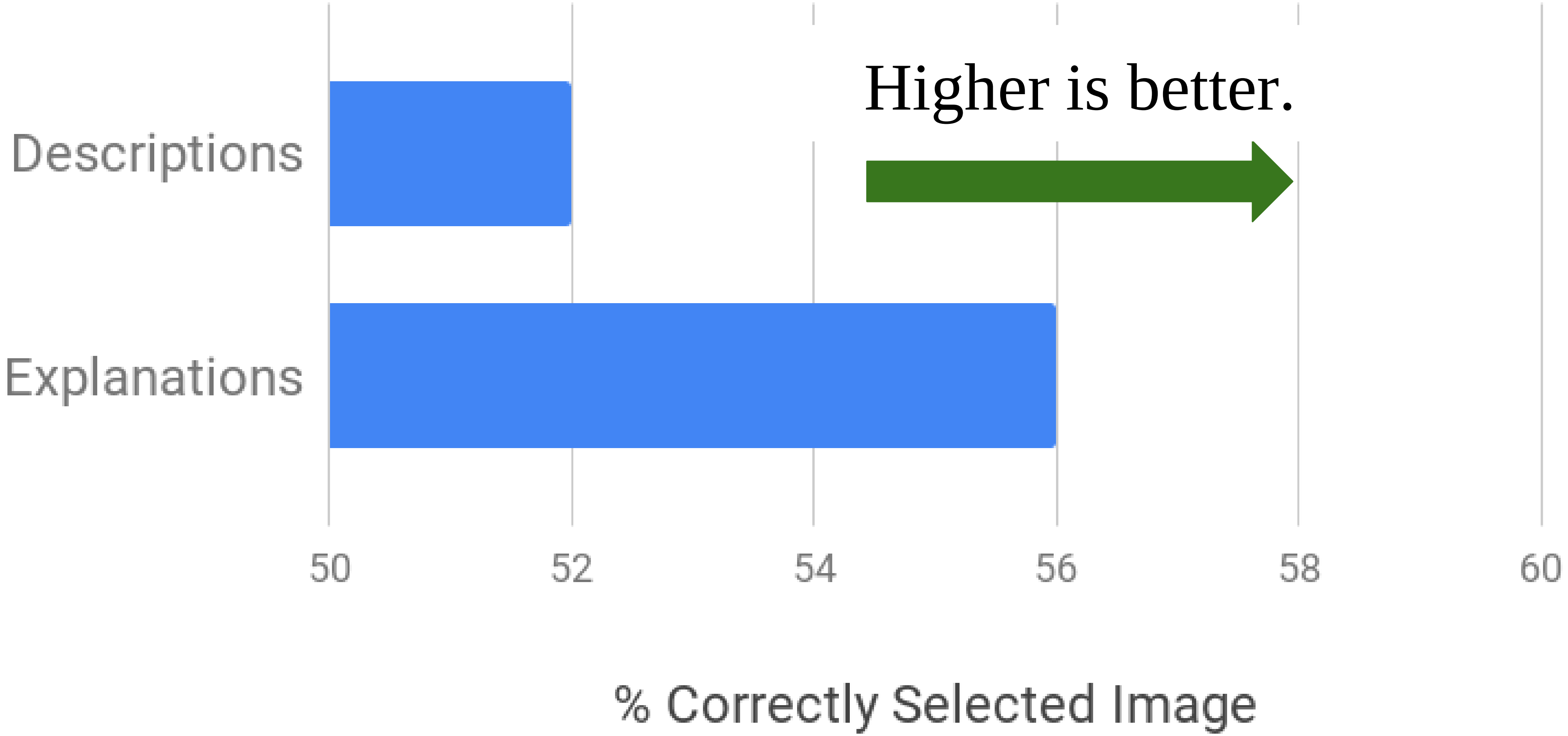
… this is a black bird with a white nape and a large black beak.

Evaluating Explanations

Choose the image which most closely matches the following text:

… this is a black bird with a white nape and a large black beak.

Which model is best for discriminating between images?



Which of the following is the best explanation for why this bird is a White Necked Raven?

A) This is a ***White Necked Raven*** because this bird is nearly all black with a short pointy bill.
B) This is a ***White Necked Raven*** because this is a black bird with a white nape and a large black beak.

Which of the following is the best explanation for why this bird is a White Necked Raven?

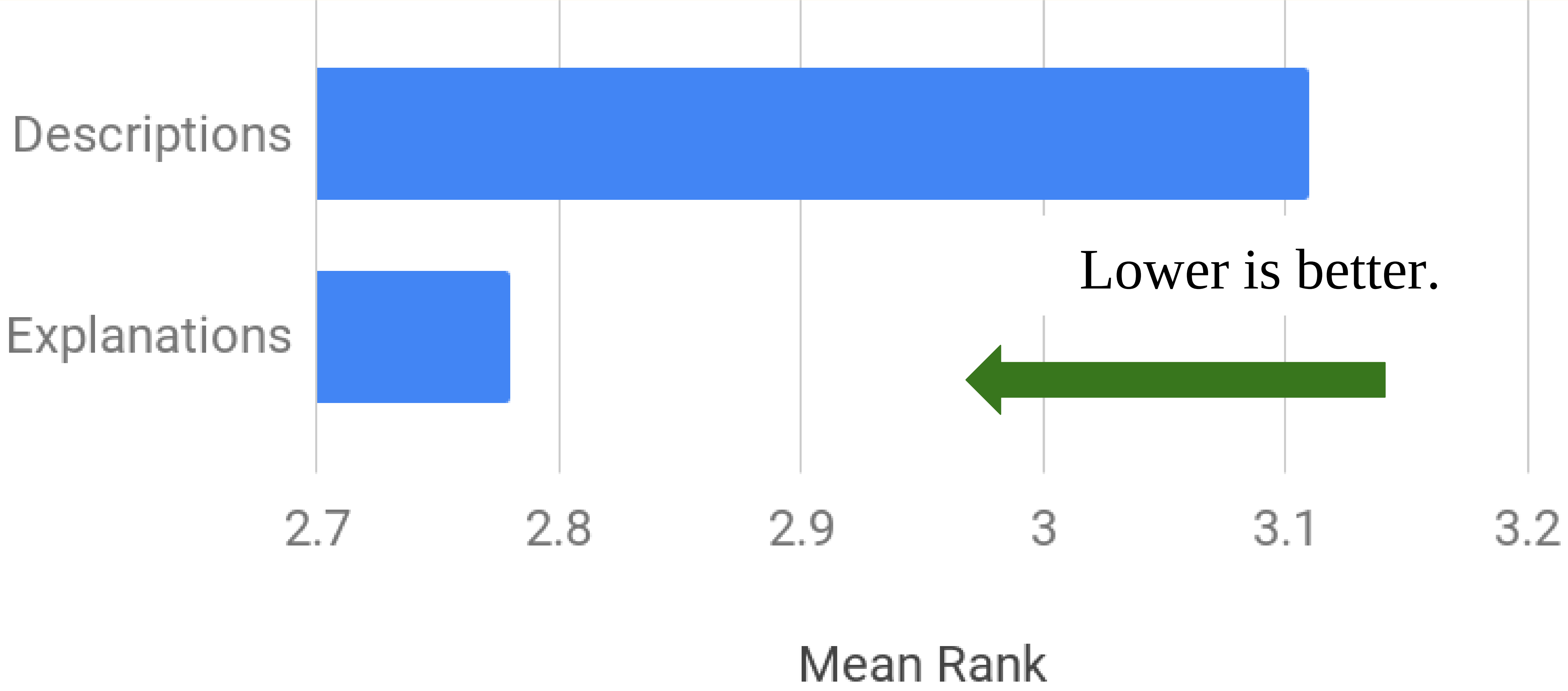
A) This is a ***White Necked Raven*** because this bird is nearly all black with a short pointy bill.
B) This is a ***White Necked Raven*** because this is a black bird with a white nape and a large black beak.

Which of the following is the best explanation for why this bird is a White Necked Raven?

A) This is a ***White Necked Raven*** because this bird is nearly all black with a short pointy bill.
B) This is a ***White Necked Raven*** because this is a black bird with a white nape and a large black beak.

Need bird watchers!

Which explanations do bird watchers prefer?



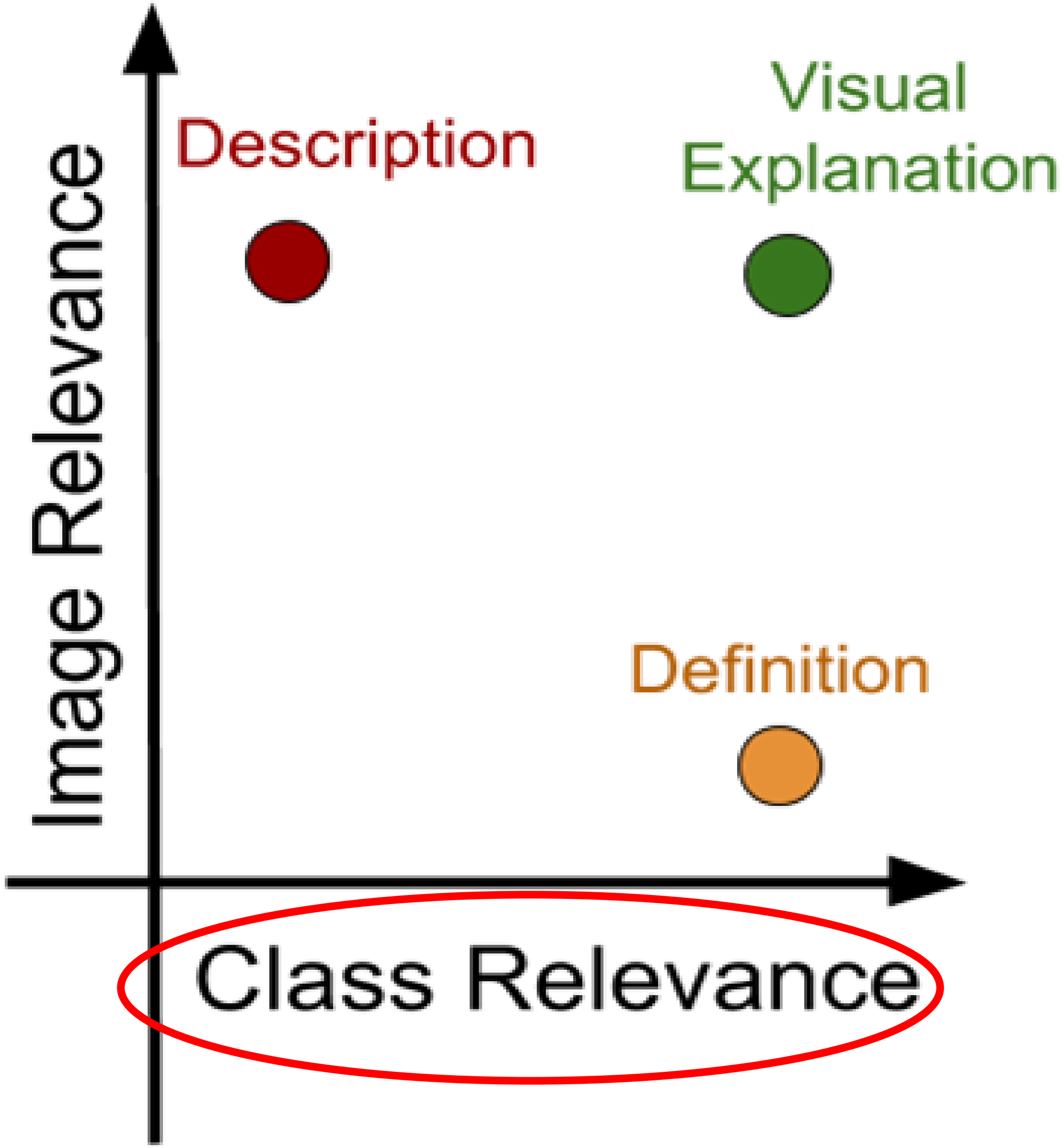


Image Relevance
Description
Visual
Explanation
Definition
Class Relevance

This is a ***mallard*** because this is a brown and white bird with a green head and a yellow beak.

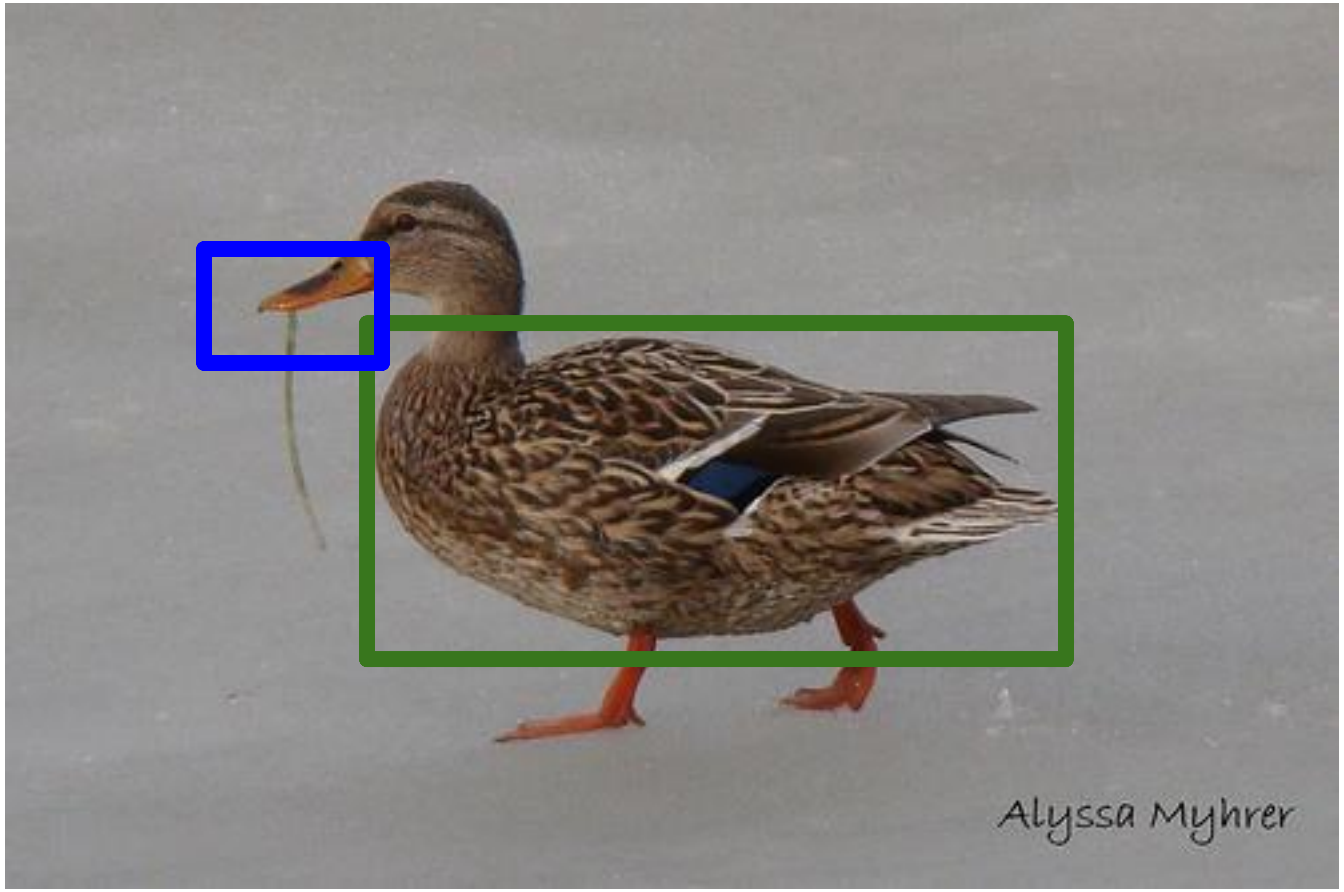

Alyssa Myhrer

This is a ***mallard*** because this is a brown and white bird with a green head and a yellow bill.

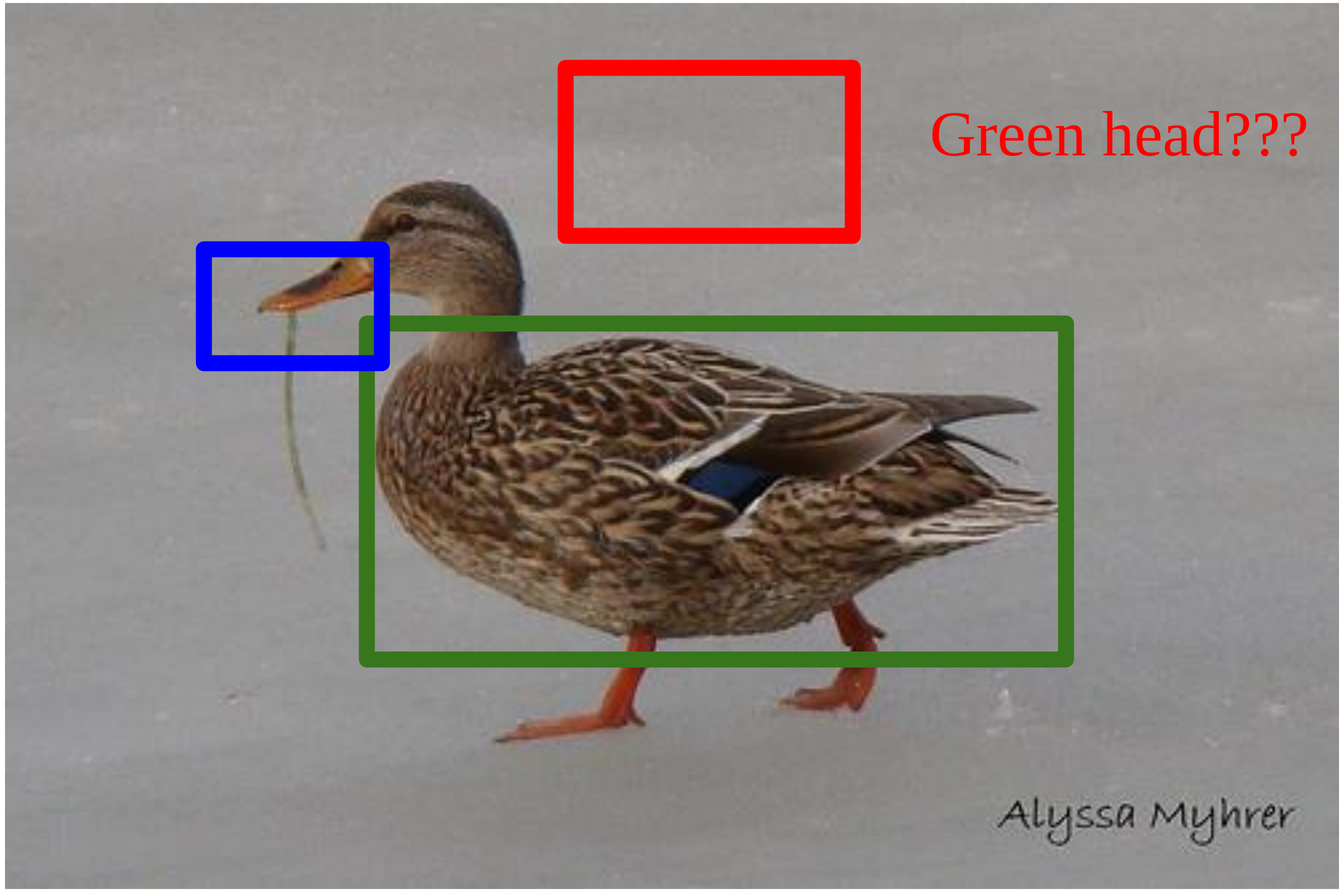
~~This is a ***mallard*** because this is a brown and white bird with a green head and a yellow bill.~~

This is a ***mallard*** because this bird has a brown head, orange feet, and a flat bill.

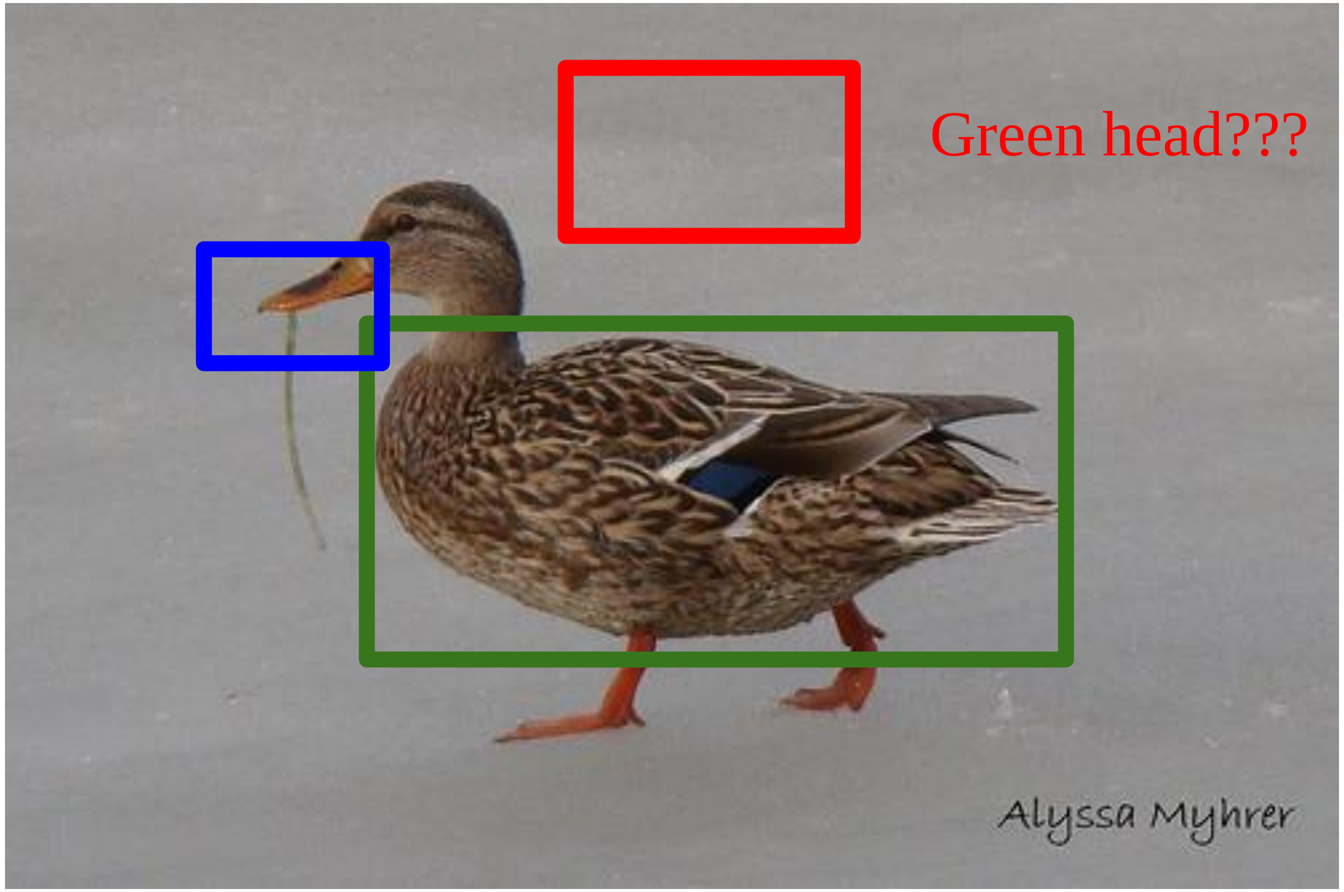
Intuition: Only output explanations which are grounded in visual evidence.



This is a ***mallard*** because this is a ***brown and white bird*** with a green head and a ***yellow bill***.

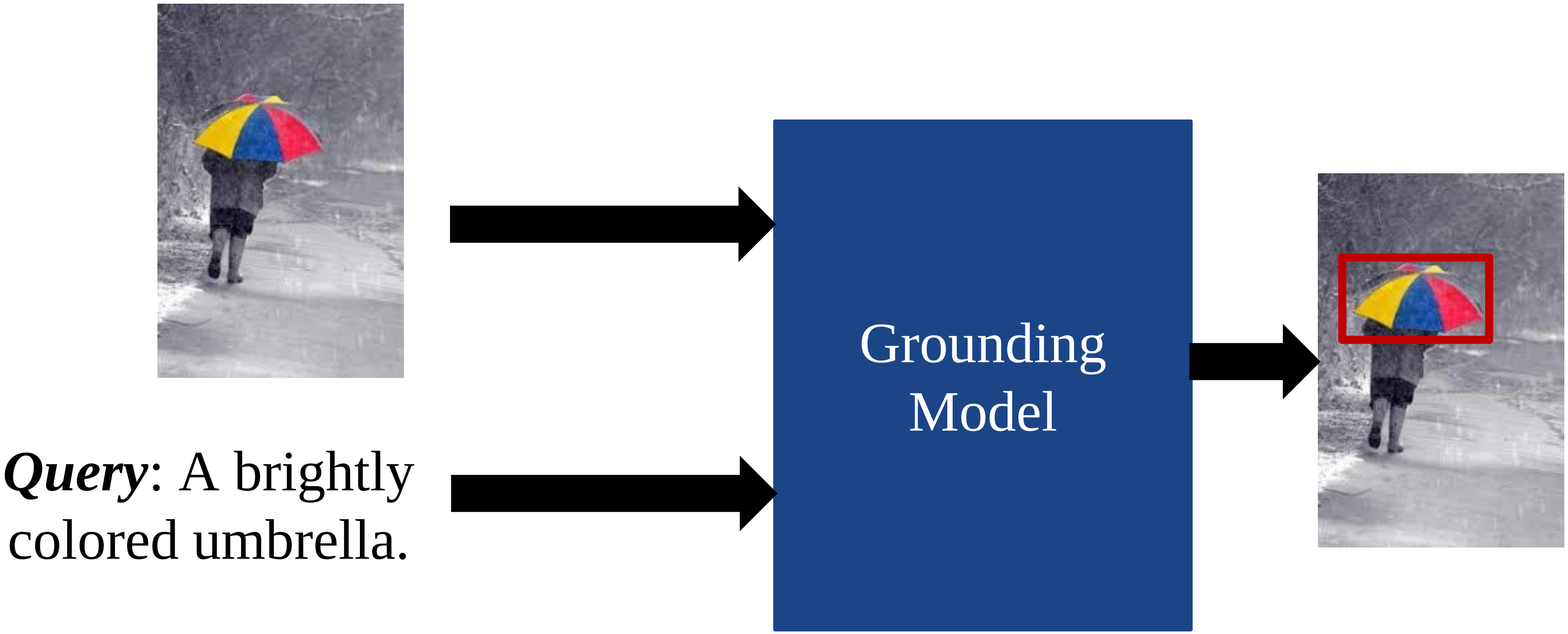


This is a ***mallard*** because this is a ***brown and white bird*** with a ***green head*** and a ***yellow bill***.

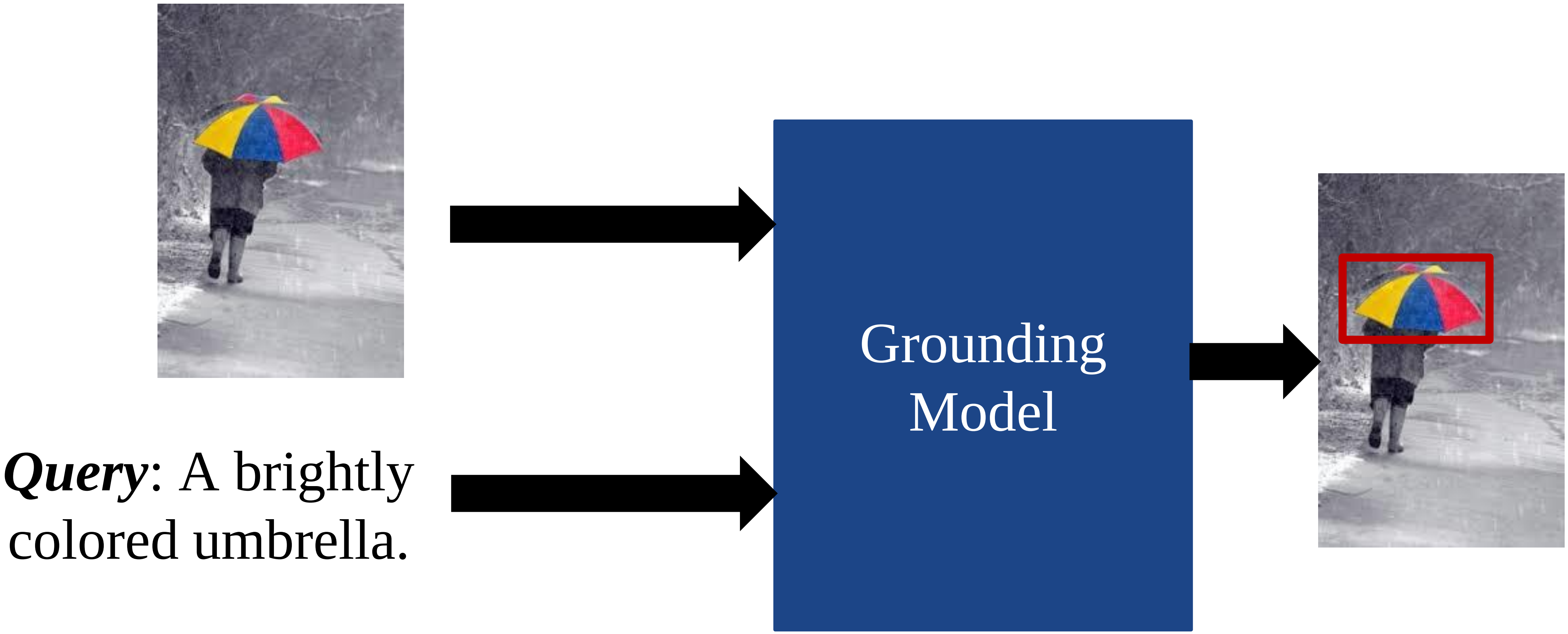


This is a ***mallard*** because this is a ***brown and white bird*** with a ***green head*** and a ***yellow bill***.

- Call “brown and white bird”, “green head”, and “yellow bill” attributes.
- Extract attributes with a noun phrase chunker.

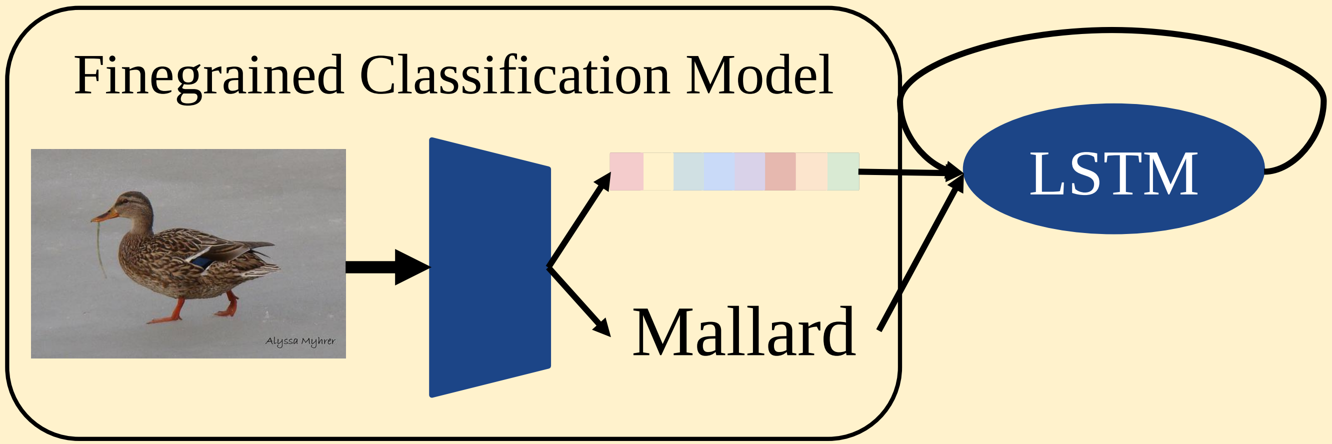


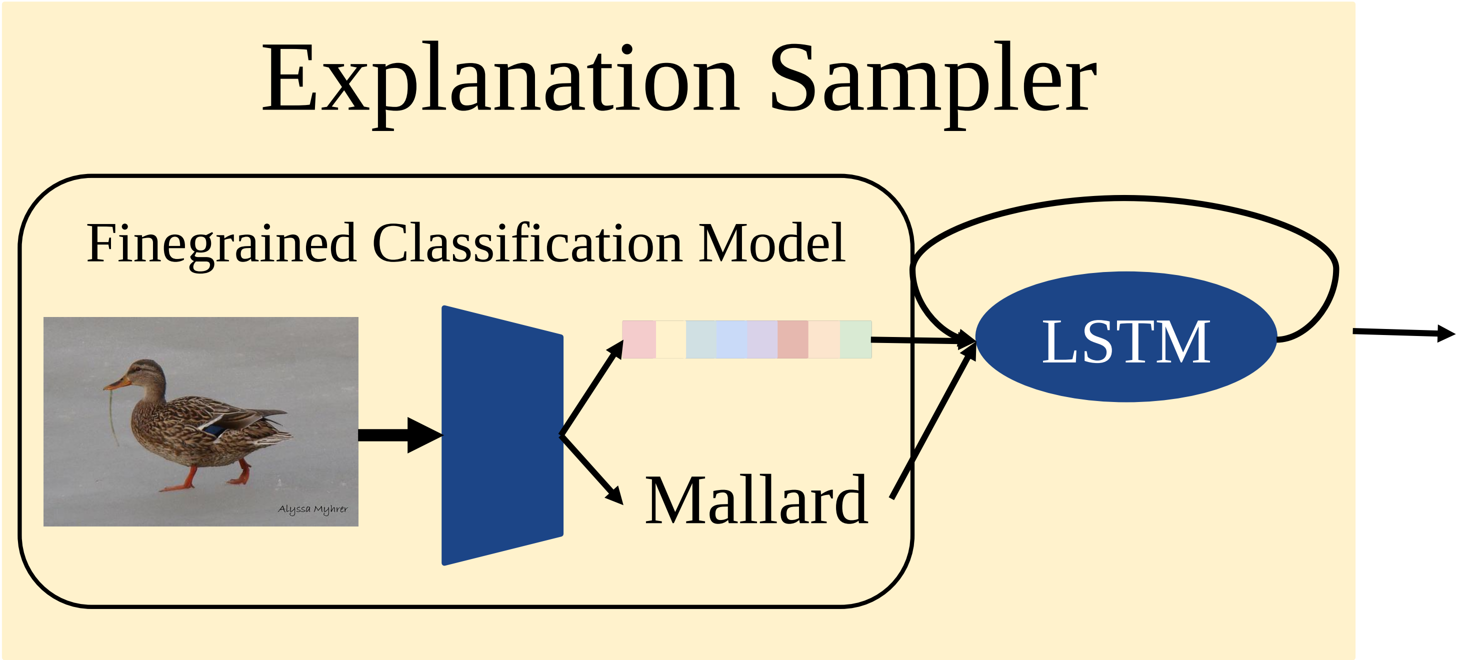
Model from:
Hu et al. *Modeling Relationships in Referential Expressions with Compositional Modular Networks*. CVPR 2017.



Model from:
Hu et al. *Modeling Relationships in Referential Expressions with Compositional Modular Networks*. CVPR 2017.
Trained with Visual Genome:
Krishna et al. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. IJCV 2016.

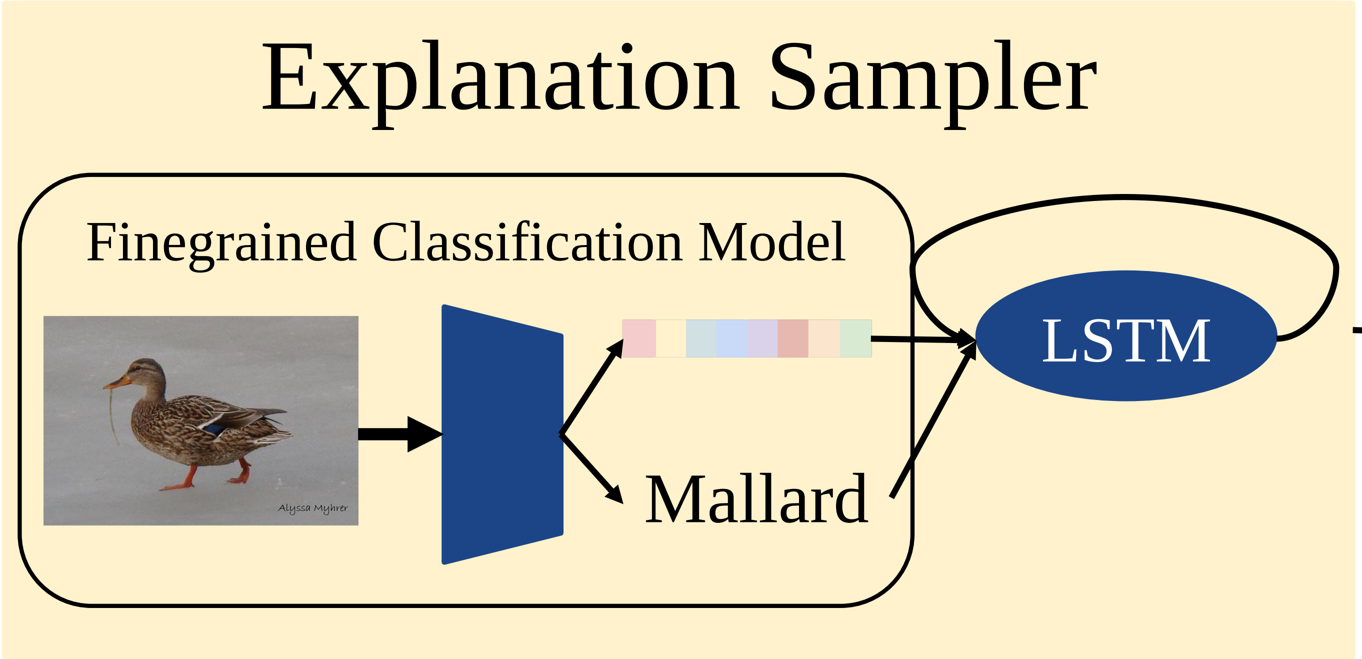
Explanation Sampler





This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

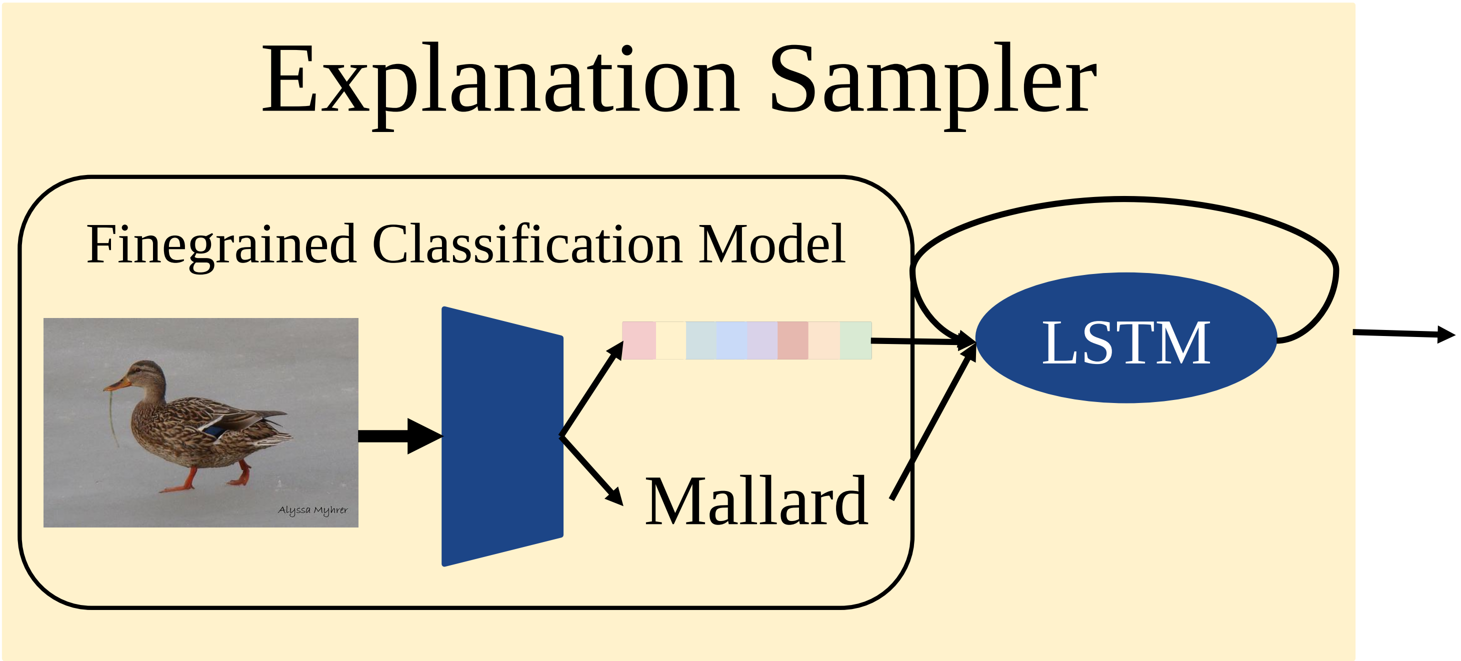


This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Generally score sentences based off *sentence fluency*:

$$S = \sum_t \log P(w_t | w_{0:t-1}, I, C)$$



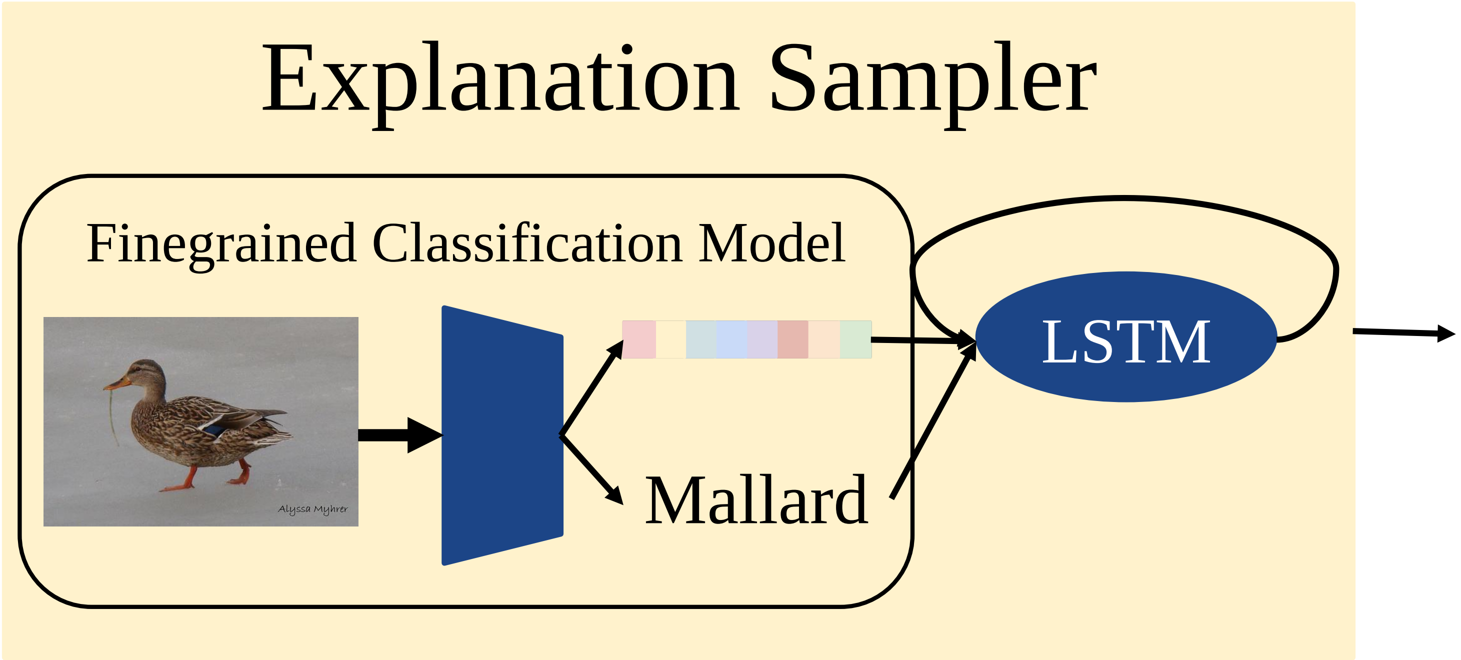
This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Baseline

Generally score sentences based off *sentence fluency*:

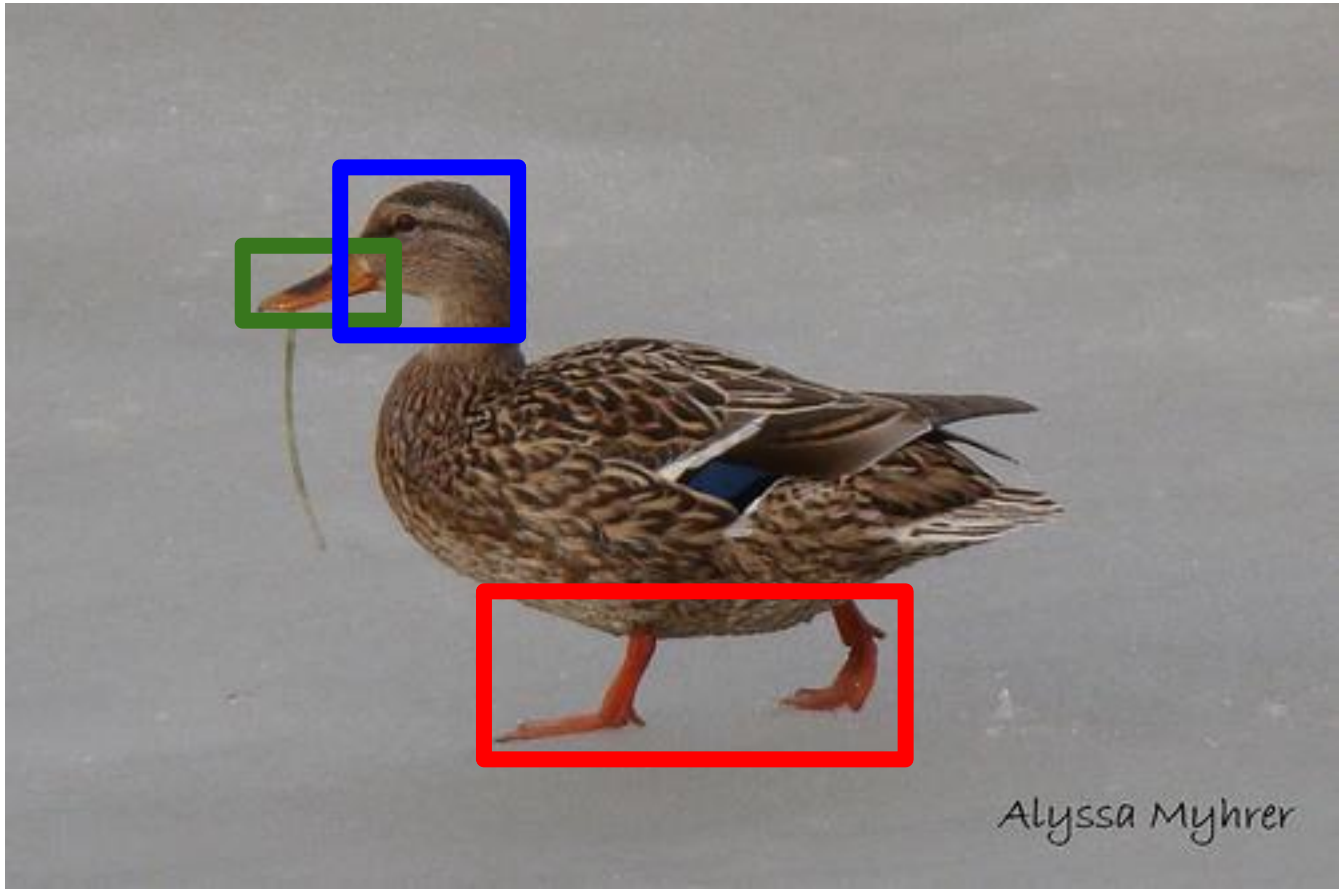
$$S = \sum_t \log P(w_t | w_{0:t-1}, I, C)$$



This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

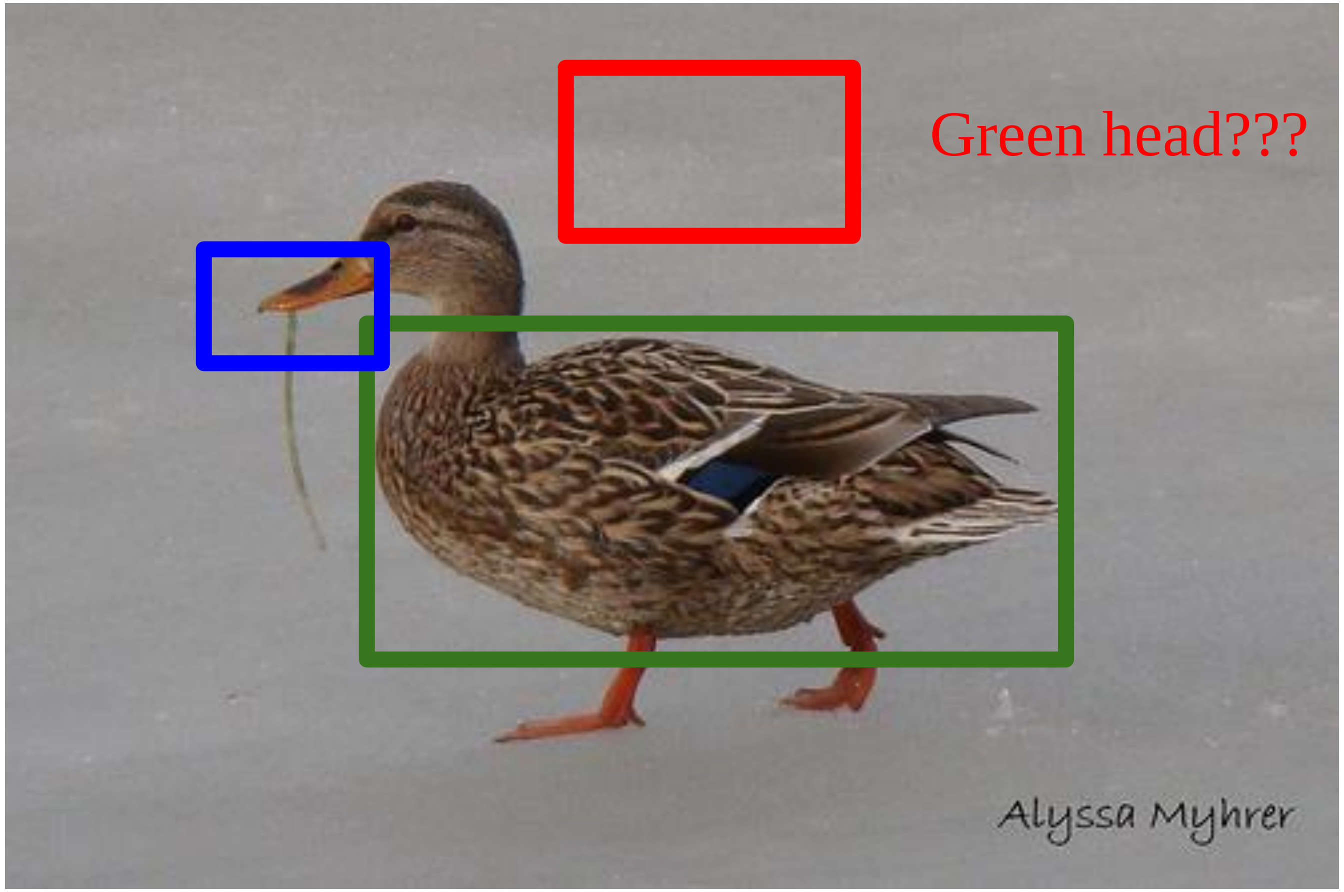
Can we score sentences on visual grounding instead?



This is a ***mallard*** because this bird has a brown head, orange feet, and a flat bill.

Score for *brown head*: 1.9
Score for *orange feet*: 2.1
Score for *flat bill*: 1.1

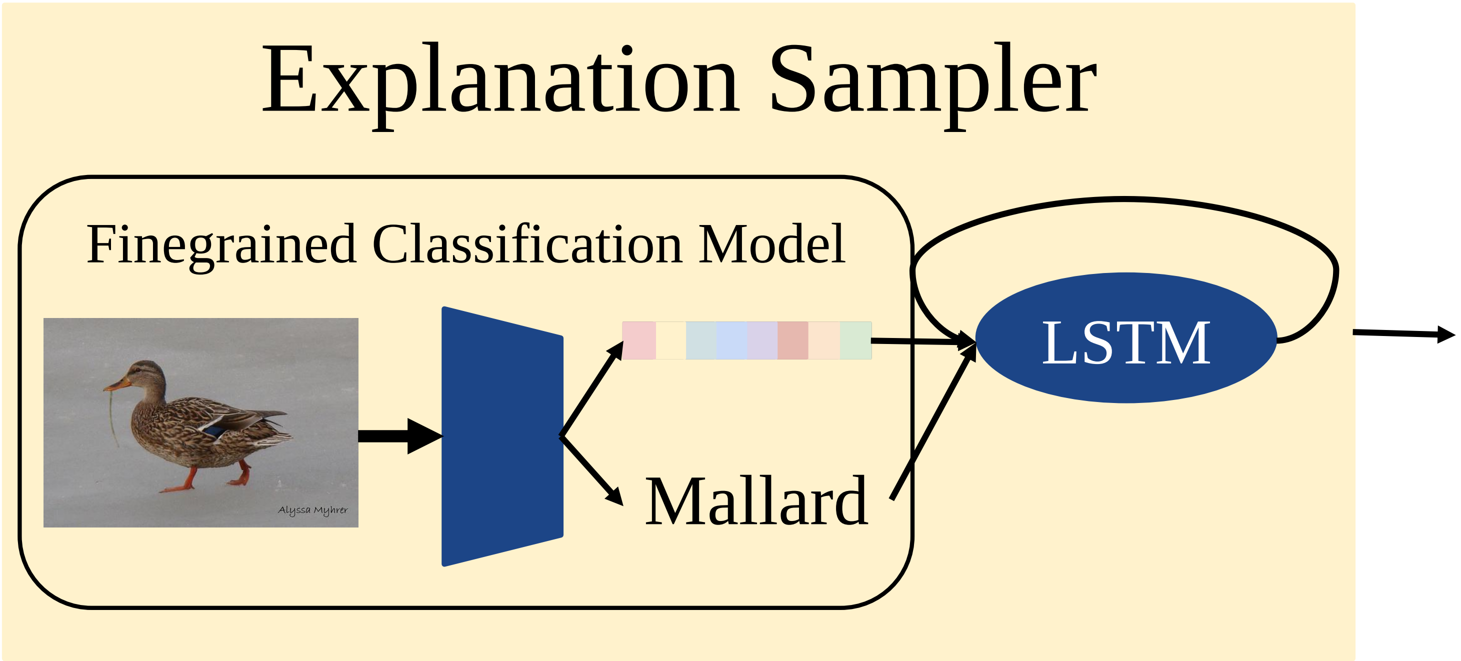
Average score high → good explanation.



This is a ***mallard*** because this is a brown and white bird with a green head and a yellow bill.

Score for *brown and white bird*: 2.2
Score for *green head*: 0.2
Score for *yellow beak*: 1.2

Average score low → bad explanation.



This is a mallard because this is a brown and white bird with a green head and a yellow bill.

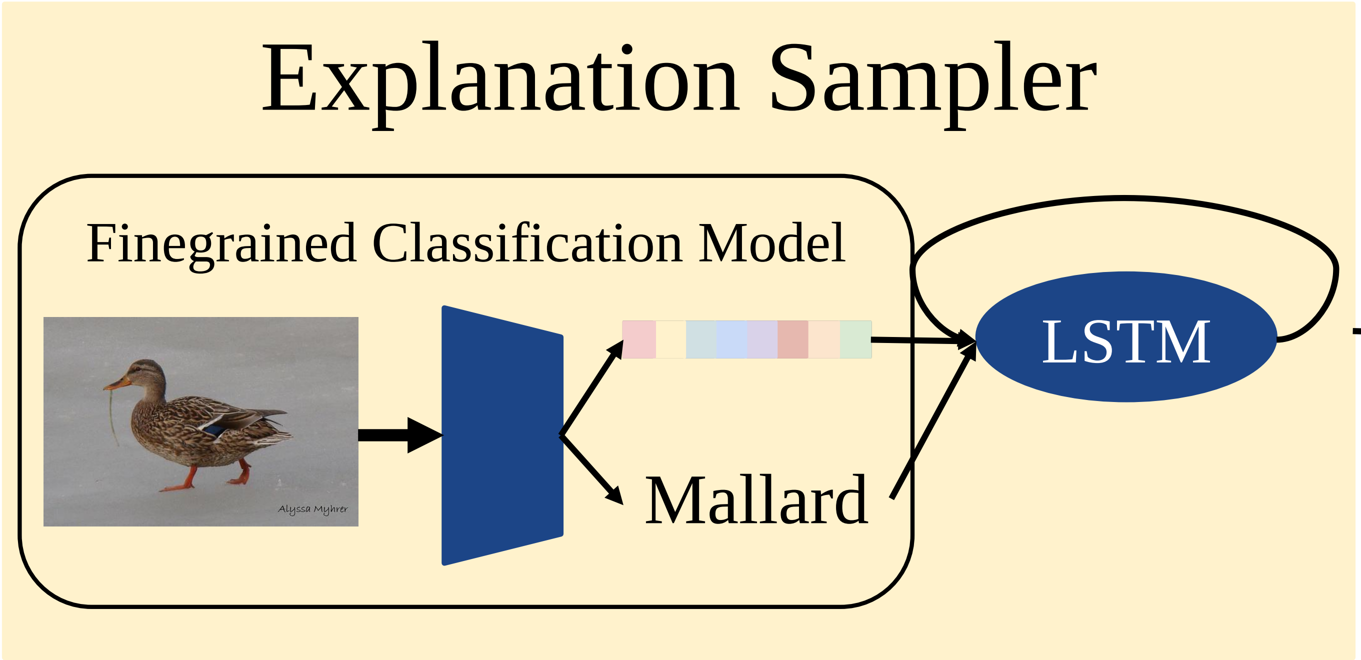
This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Score sampled sentences with visual grounding model.

A is set of attributes in explanation

$$S = \frac{1}{|A|} \sum_{a \in A} GroundingScore\,(a, I)$$

Grounding score for attribute in image.



This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Average grounding

Score sampled sentences with visual grounding model.

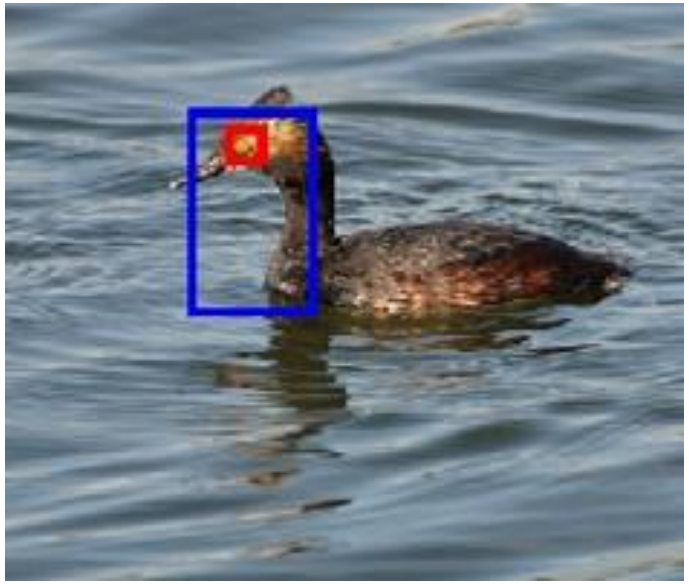
A is set of attributes in explanation

$$S = \frac{1}{|A|} \sum_{a \in A} GroundingScore\,(a, I)$$

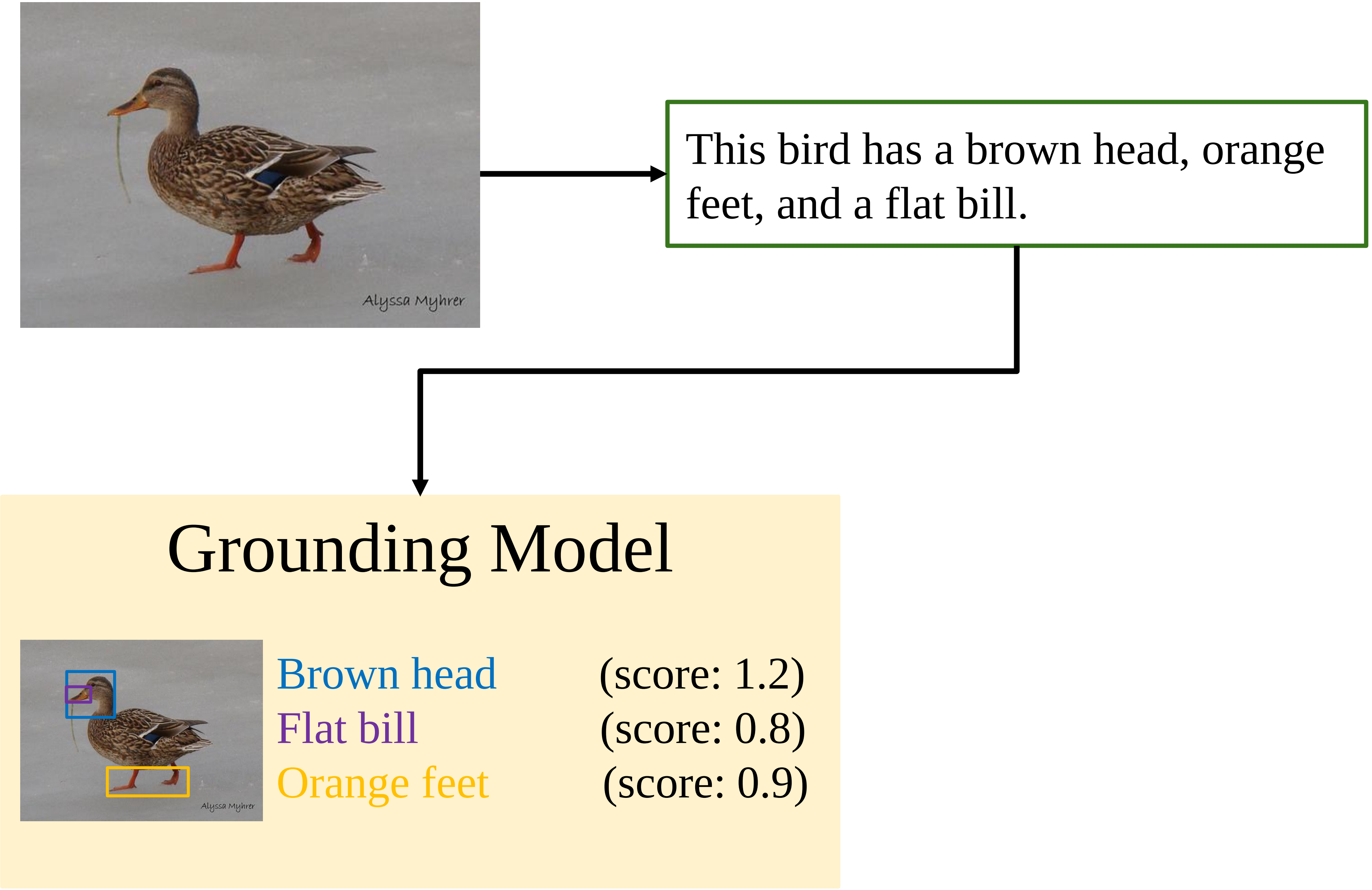
Grounding score for attribute in image.

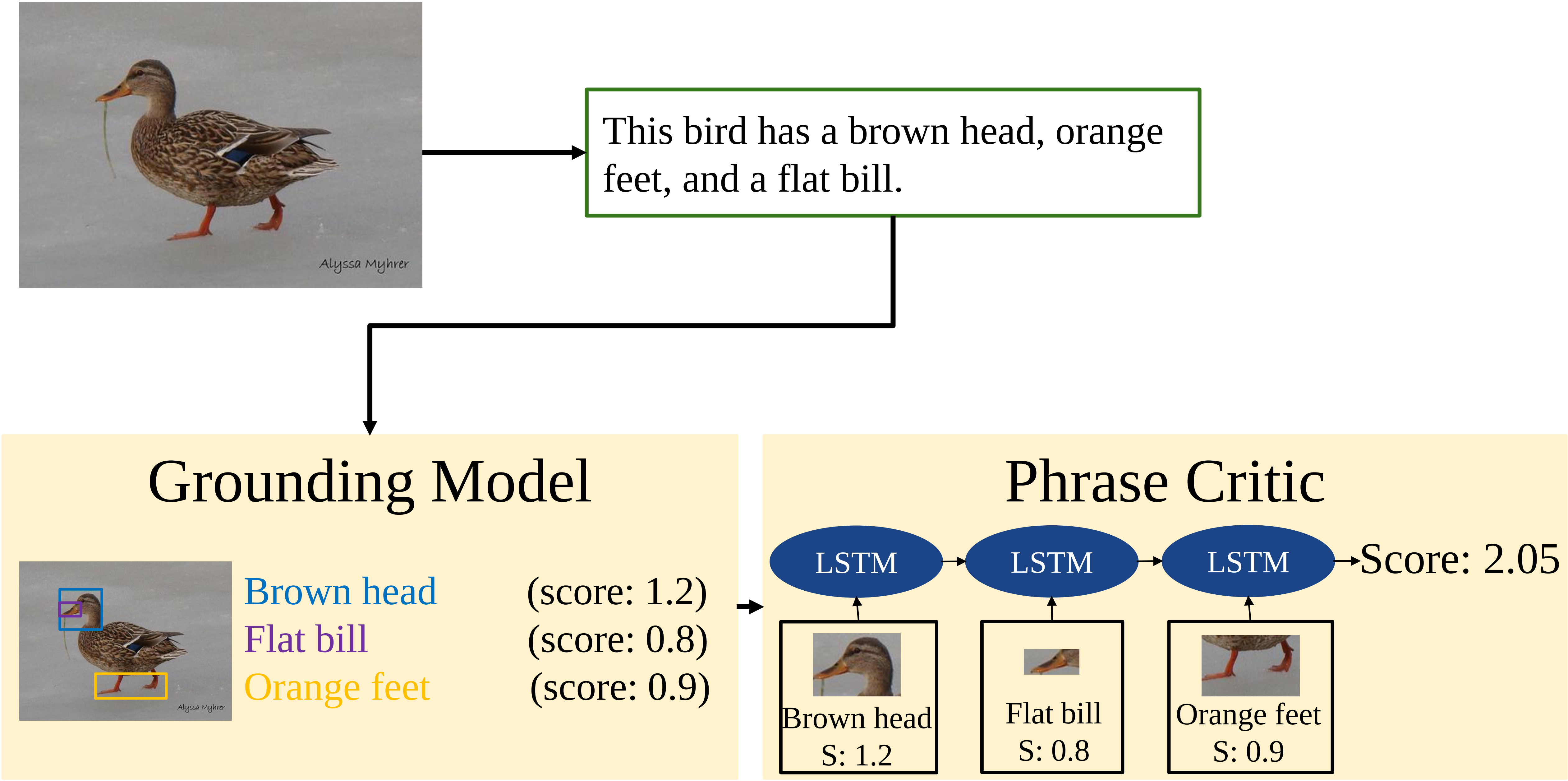
This is a **Eared Grebe** because ….

Baseline:
this is a black bird
with a long neck and
red eyes.

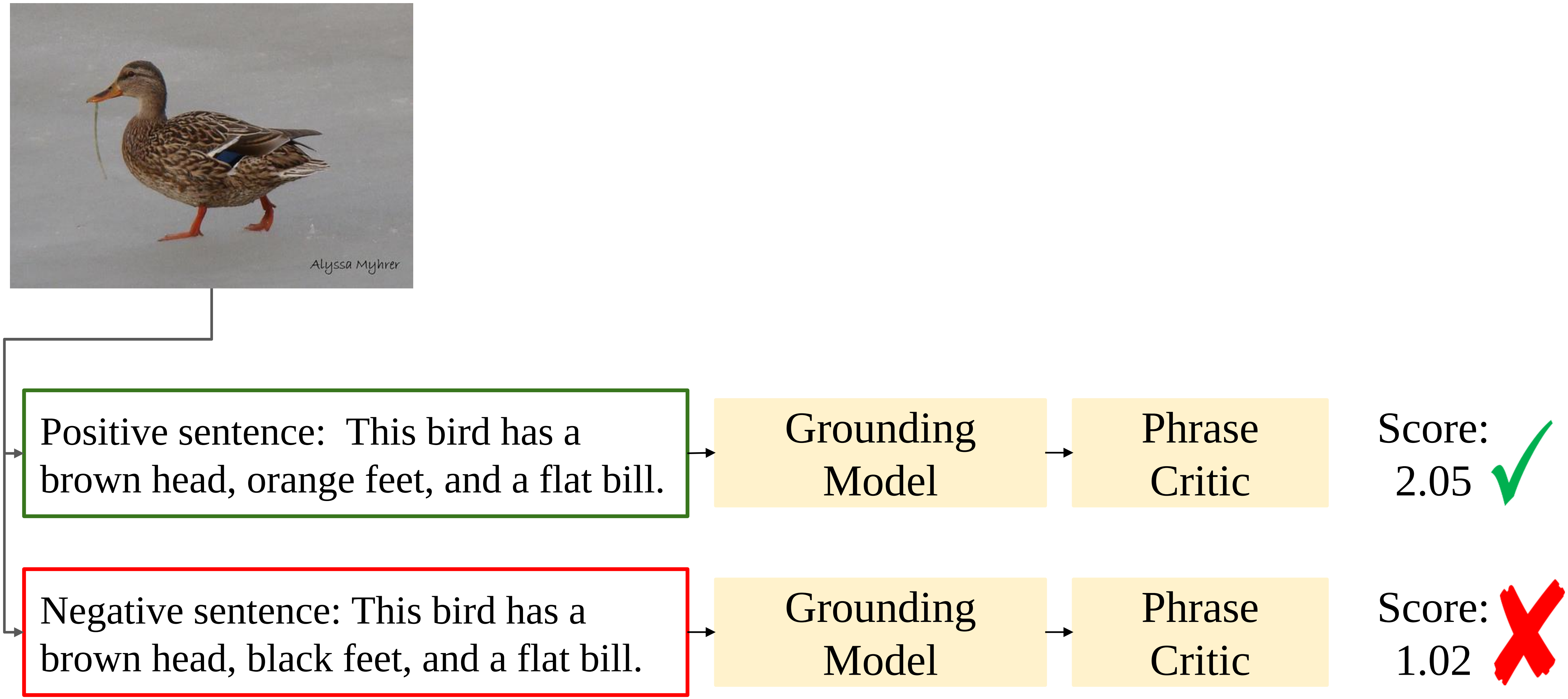


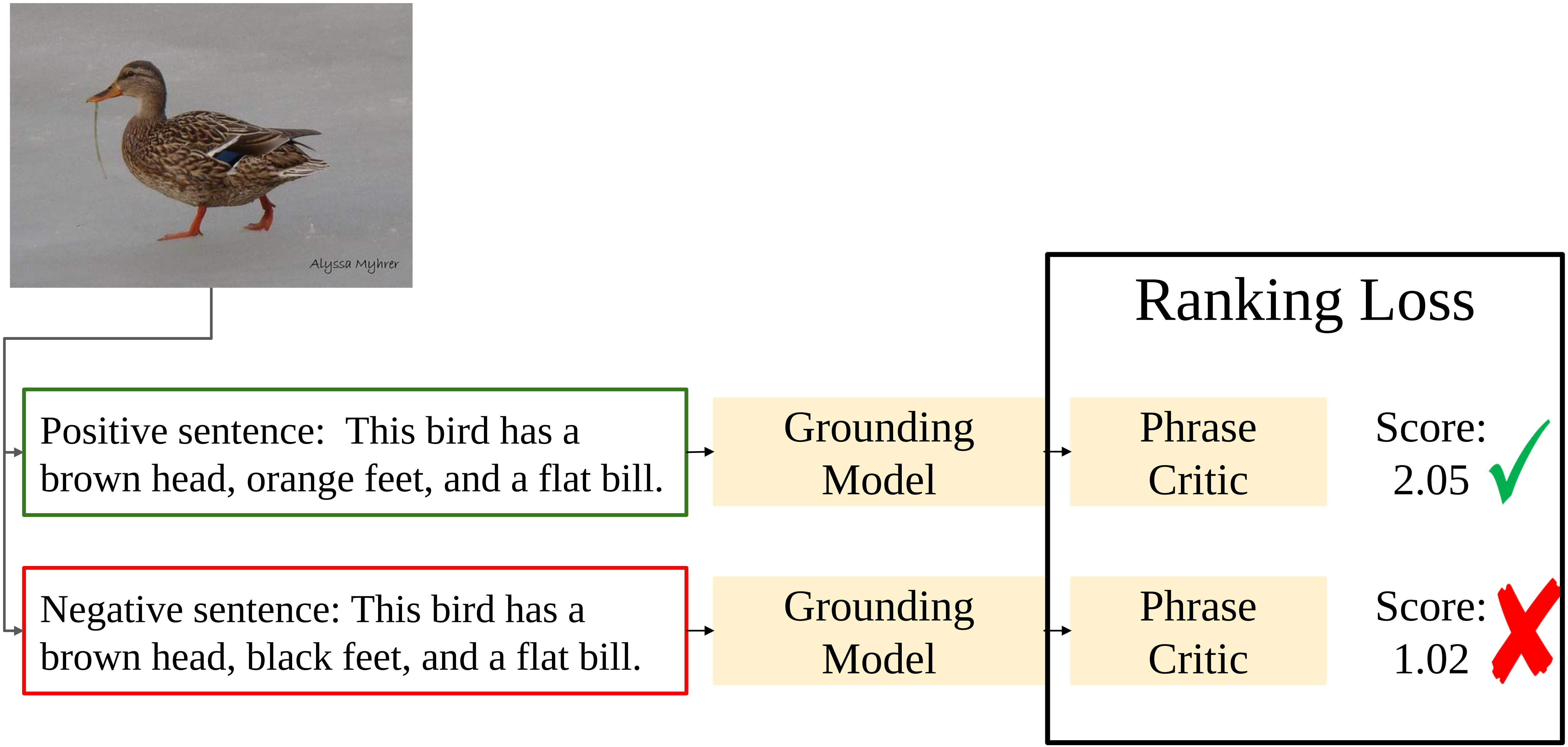
Average Grounding:
…this is a **black bird**
with a **white eye** and a
red eye.


Alyssa Myhrer
This bird has a brown head, orange feet, and a flat bill.
Grounding Model
Brown head (score: 1.2)
Flat bill (score: 0.8)
Orange feet (score: 0.9)
Alyssa Myhrer



This bird has a brown head, orange feet, and a flat bill.
Grounding Model
Brown head (score: 1.2)
Flat bill (score: 0.8)
Orange feet (score: 0.9)
Phrase Critic
LSTM
LSTM
LSTM
Score: 2.05
Brown head
S: 1.2
Flat bill
S: 0.8
Orange feet
S: 0.9
Alyssa Myhrer


Alyssa Myhrer
Positive sentence: This bird has a brown head, orange feet, and a flat bill.
Grounding Model
Phrase Critic
Score: 2.05
Negative sentence: This bird has a brown head, black feet, and a flat bill.
Grounding Model
Phrase Critic
Score: 1.02

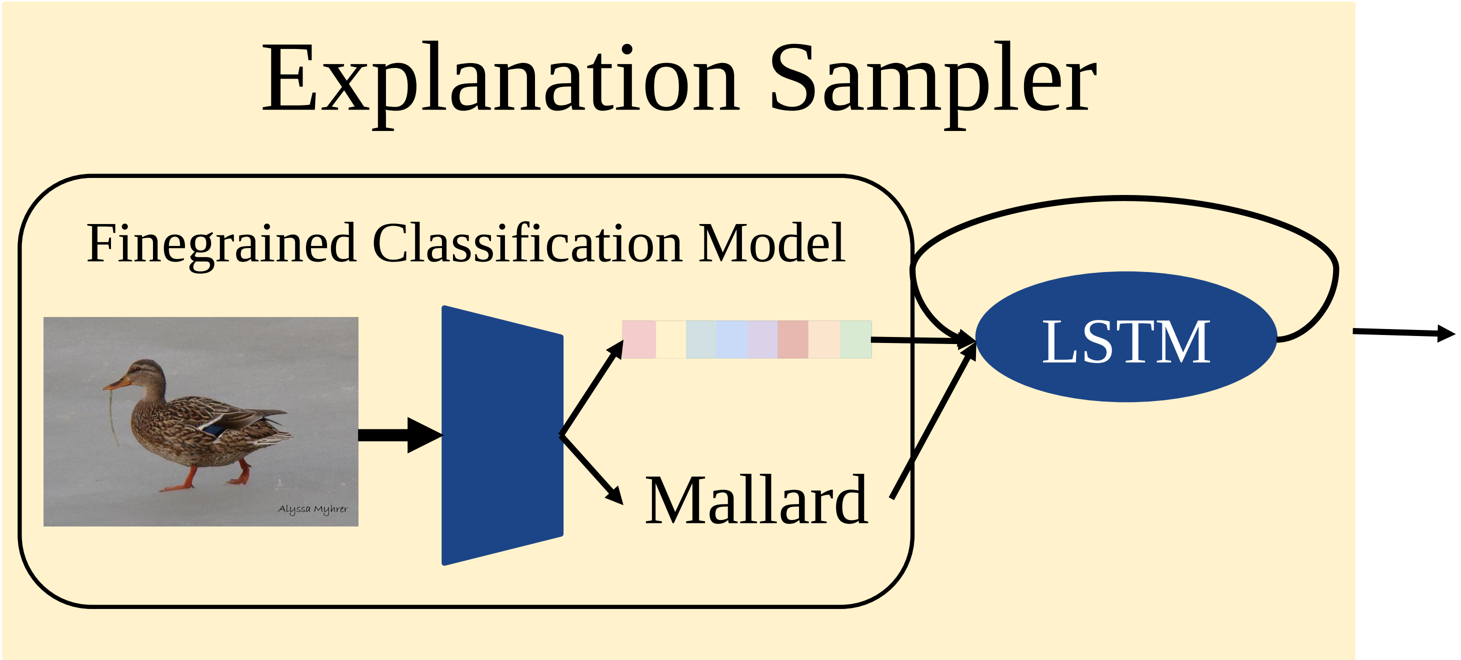

Alyssa Myhrer
Ranking Loss
Positive sentence: This bird has a brown head, orange feet, and a flat bill.
Grounding Model
Phrase Critic
Score: 2.05
Negative sentence: This bird has a brown head, black feet, and a flat bill.
Grounding Model
Phrase Critic
Score: 1.02

Positive Sentence: This bird has a brown head, orange feet and a flat bill.

Positive Sentence: This bird has a brown head, **orange** feet and a flat bill.

Positive Sentence: This bird has a brown head, **orange** feet and a flat bill.

Negative Sentence: This bird has a brown head, **black** feet and a flat bill.



This is a mallard because this is a brown and white bird with a green head and a yellow bill.

This is a mallard because this bird has a brown head, orange feet, and a flat bill.

Score sampled sentences with phrase critic.

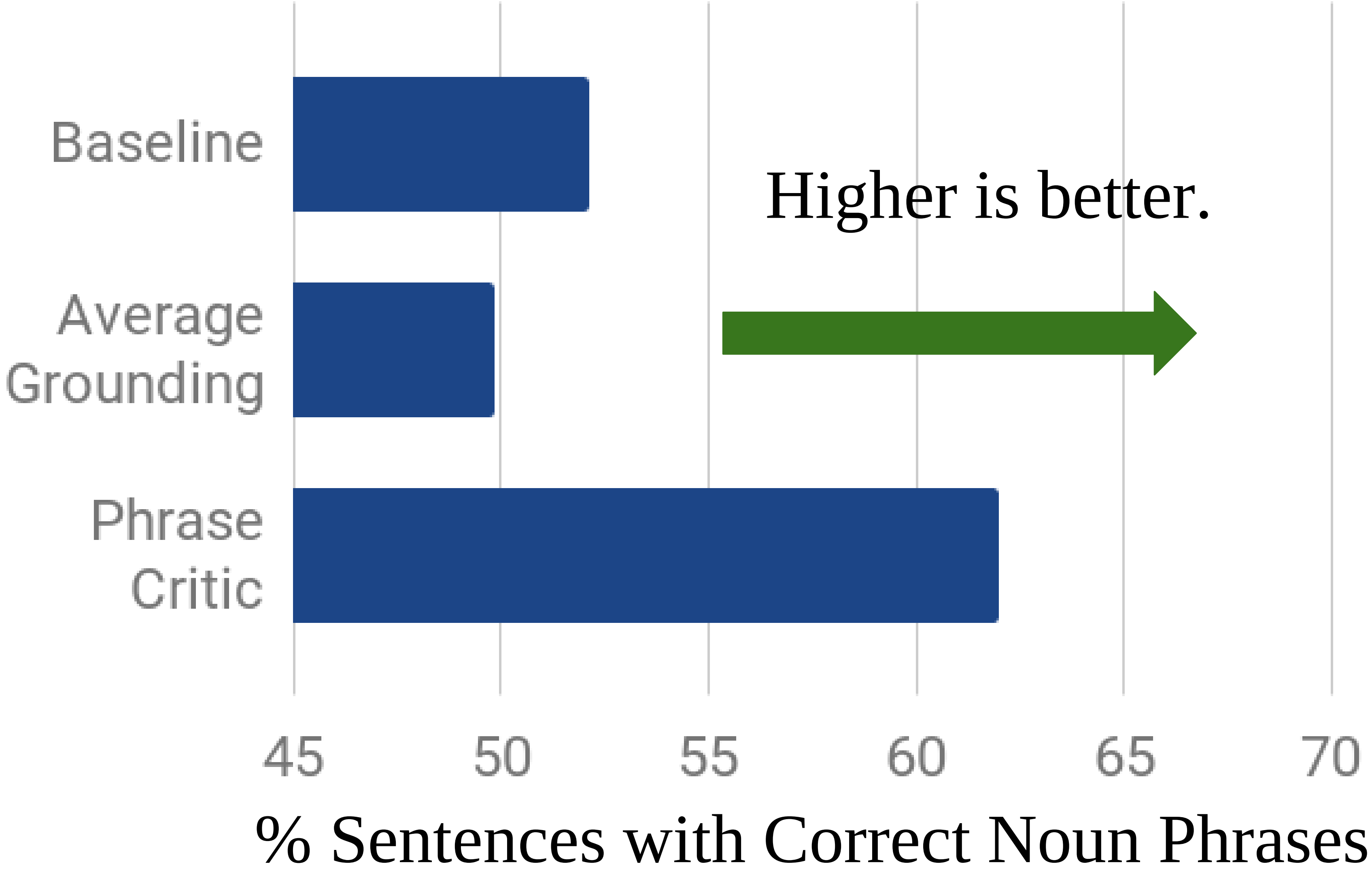
$$S = PhraseCritic(A, I)$$

Extracted noun phrase from explanation: brown and white bird, green head, yellow bill.

Does this bird have a ***green head***?

Does this bird have a ***green head***?

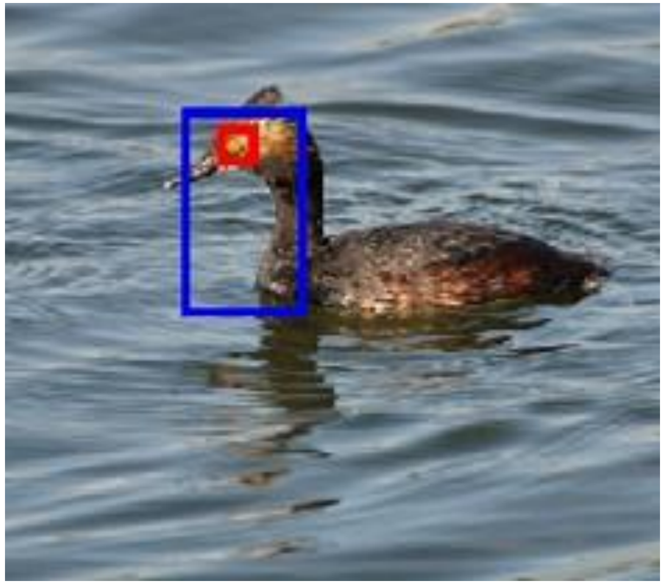
Are grounded explanations more image relevant?



This is a **Eared Grebe** because ….

Baseline:
this is a black bird
with a long neck
and red eyes

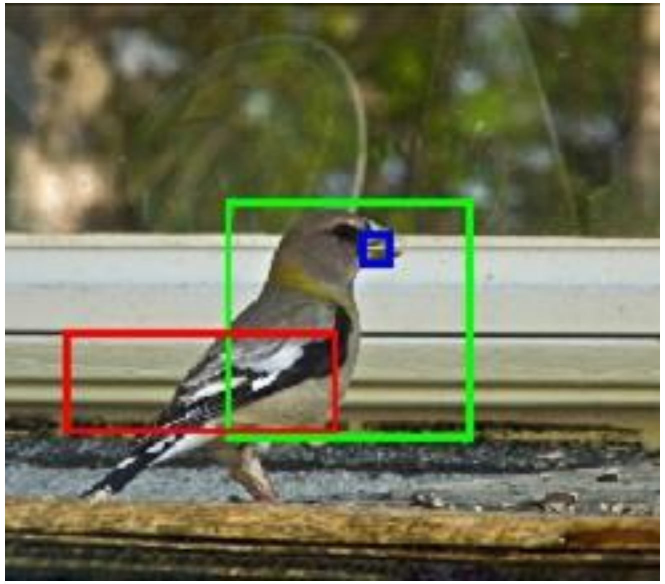
Average grounding:
this is a **black bird**
with a **white eye**
and a **red eye**.



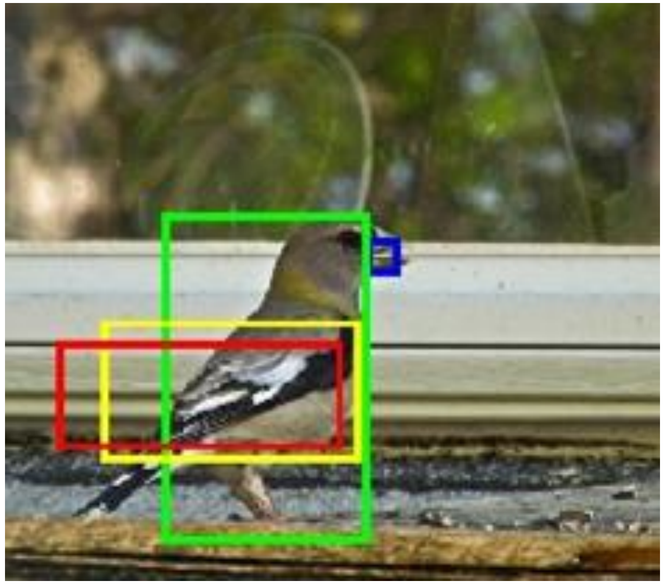
Phrase critic:
this bird has a **long neck** and **bright orange eyes**.

This is a **Evening Grosbeak** because ….

Baseline:
this is a yellow bird with a black and white wing and a yellow beak.



Average grounding:
this is a **white bird** with a **brown and black wing** and a **yellow beak**.



Phrase critic:
this is is a **small brown bird** with a **white and black wing** and a **yellow beak**.


What type of bird is this?
It is a **Cardinal** because it is a **red bird** with a **red beak** and a **black face**


Why isn’t it a ***Scarlet Tanager***?
It isn’t a ***Scarlet Tanager*** because it doesn’t have black wings.

Pipeline:

Why isn’t this a Scarlet Tanager?

Pipeline:

Why isn’t this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a ***red bird*** with ***black wings***.
This ***red bird*** has a ***pointy beak*** and ***black eyes***.
...

Pipeline:

Why isn't this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a ***red bird*** with ***black wings***.
This ***red bird*** has a ***pointy beak*** and ***black eyes***.
...

Ground Scarlet Tanager evidence:

Red bird: grounded
Pointy beak: grounded
…
Black wings: Not grounded!

Pipeline:

Why isn't this a Scarlet Tanager?

Predict evidence for Scarlet Tanager:

This is a ***red bird*** with ***black wings***.
This ***red bird*** has a ***pointy beak*** and ***black eyes***.
...

Ground Scarlet Tanager evidence:

Red bird: grounded
Pointy beak: grounded
…
Black wings: Not grounded!

Construct sentence:

This is not a ***Scarlet Tanager*** because it does not have ***black wings***.

Why is this a ***Blue Winged Warbler*** and not a ***Common Yellowthroat***?

Blue Winged Warbler

Common Yellowthroat

Explanation: This is a ***Blue Winged Warbler*** because this is a yellow bird with a black wing and a black pointy beak.

This is not a ***Common Yellowthroat*** because it does not have a black face.

Are Explanations Helpful to Humans?

Are Explanations Helpful to Humans?

The AI justified its prediction with the following evidence: this is a brown and black spotted bird with a white belly. Do you think you would accept the AIs prediction?

- ◯ Accept prediction
- ◯ Do not accept prediction

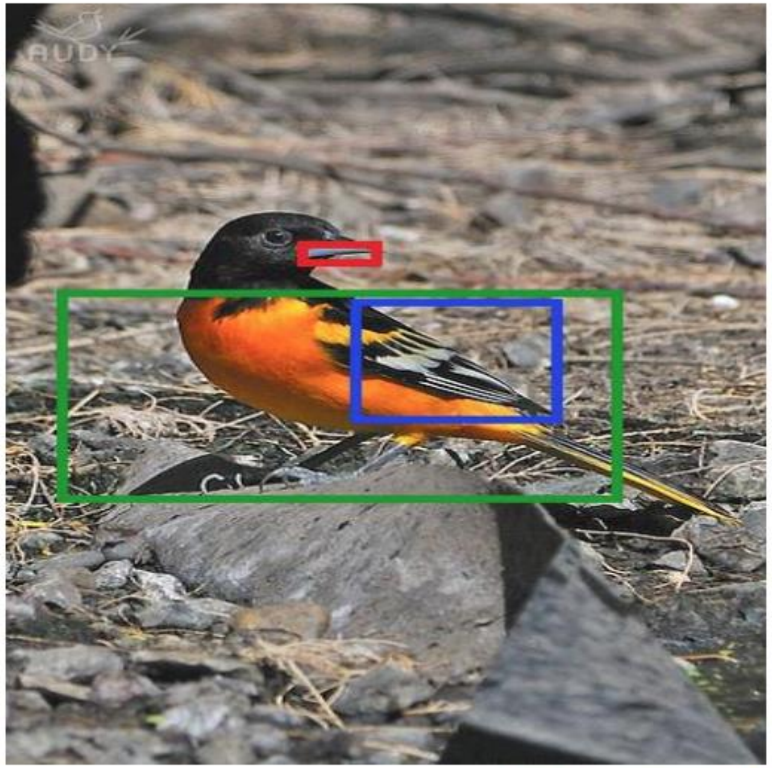
Are Explanations Helpful to Humans?

The AI is ***wrong***; you ***should not*** accept the prediction.

The AI justified its prediction with the following evidence: this is a brown and black spotted bird with a white belly. Do you think you would accept the AIs prediction?

- Accept prediction
- Do not accept prediction

Are Explanations Helpful to Humans?

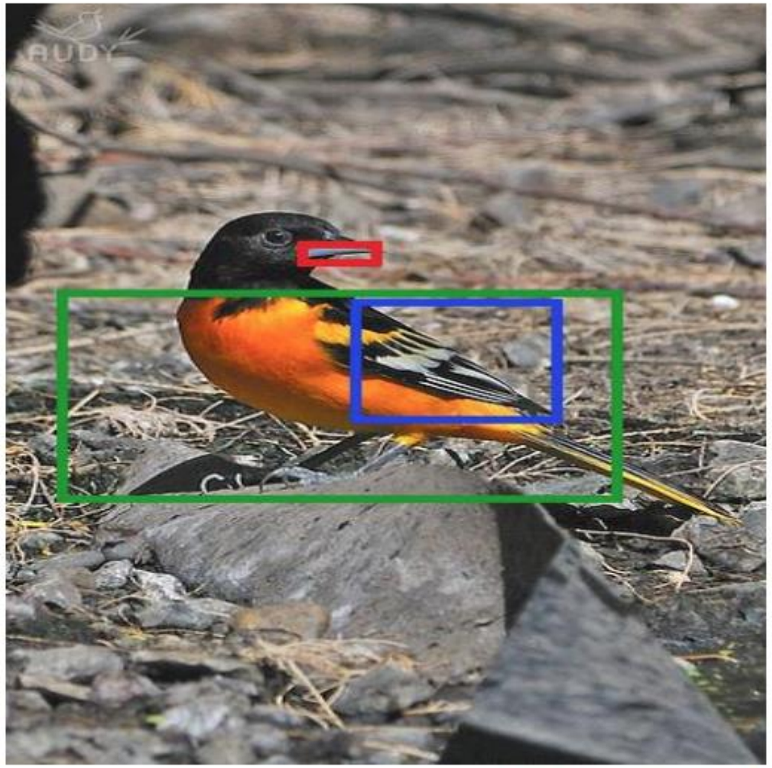


The AI justified its prediction with the following evidence: this is a small orange bird with a black wing and a small black beak. Do you think you would accept the AIs prediction?

- Accept prediction
- Do not accept prediction

Are Explanations Helpful to Humans?

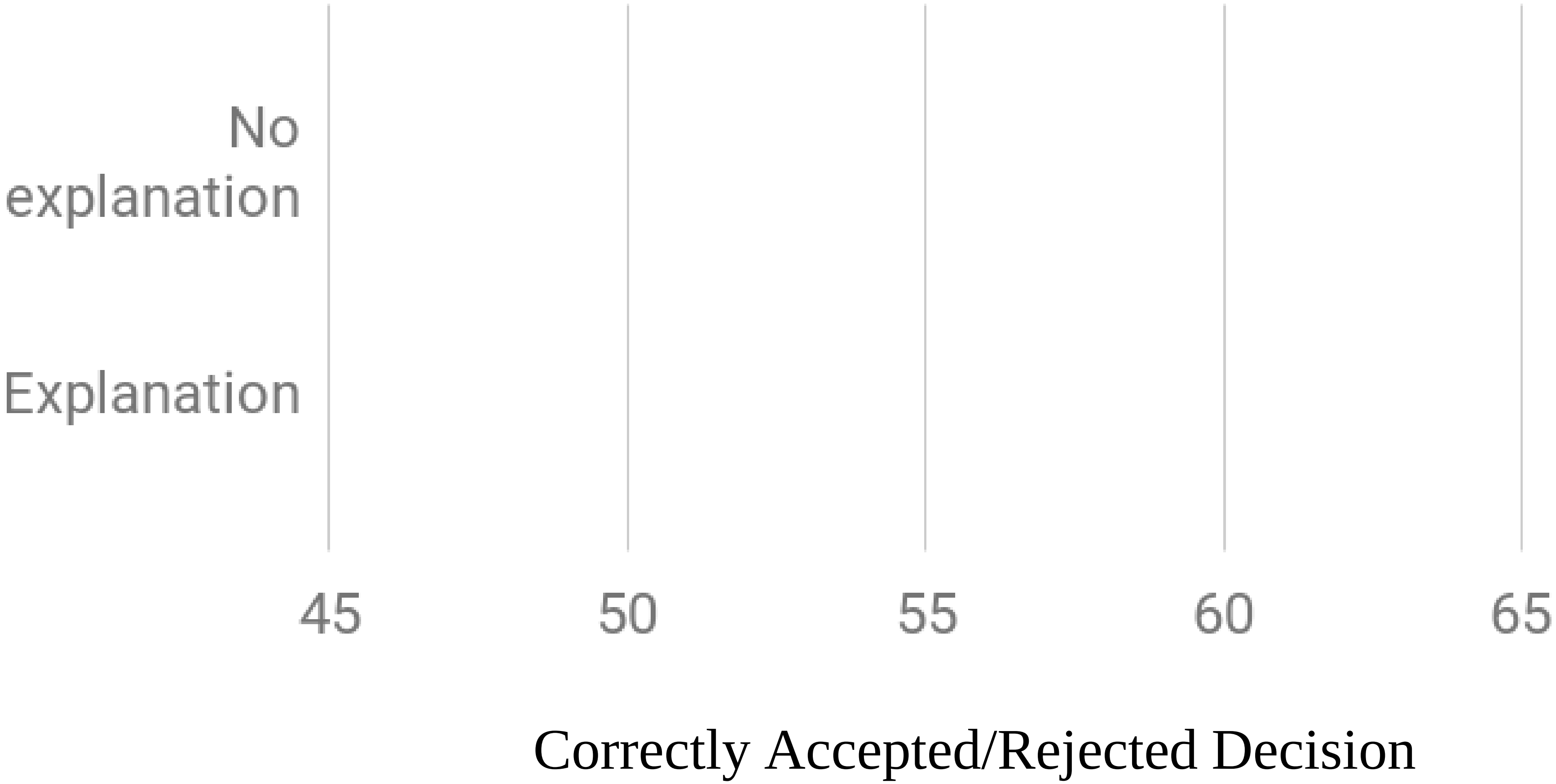
The AI is ***correct***; you ***should*** accept the prediction.



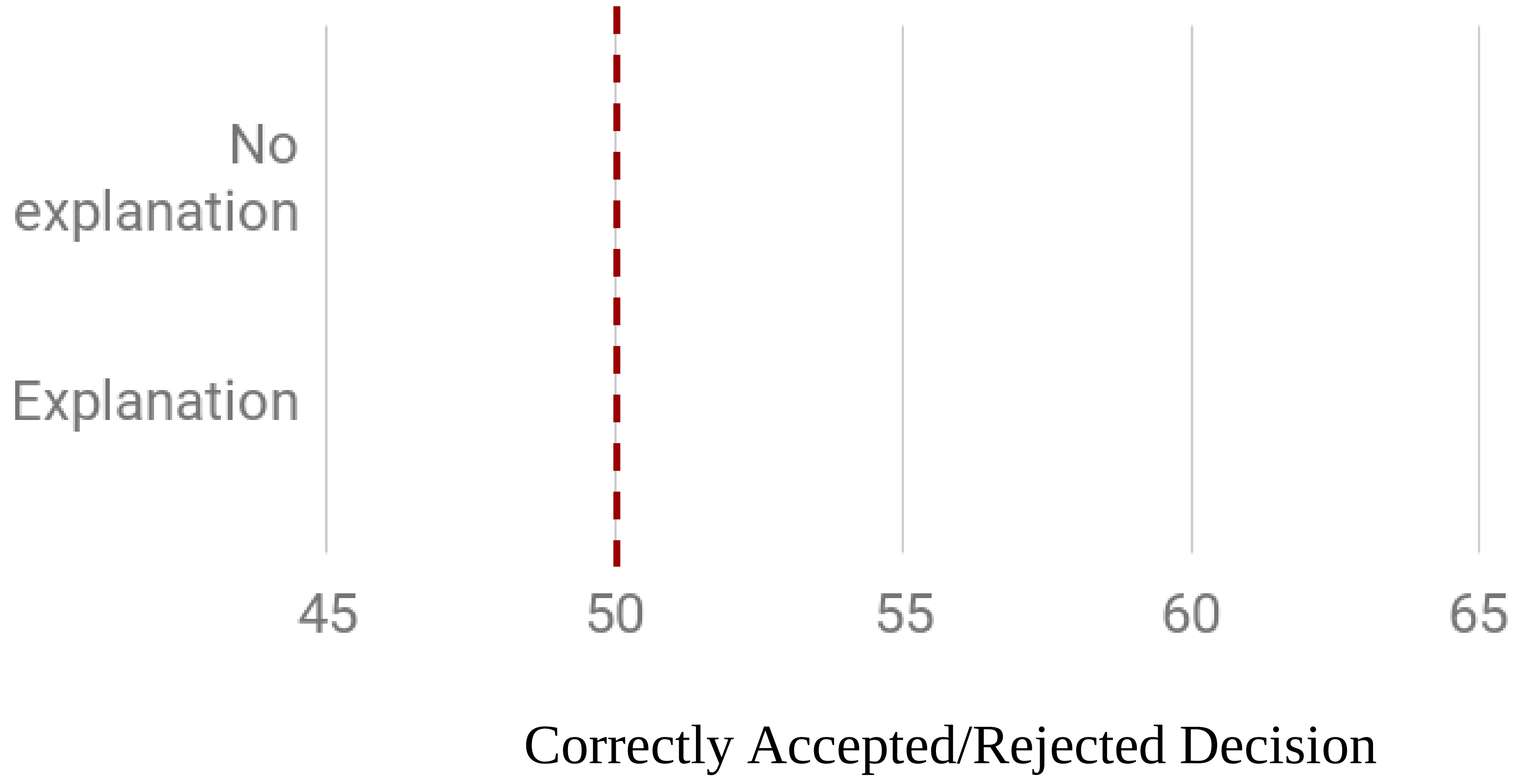
The AI justified its prediction with the following evidence: this is a small orange bird with a black wing and a small black beak. Do you think you would accept the AIs prediction?

- Accept prediction
- Do not accept prediction

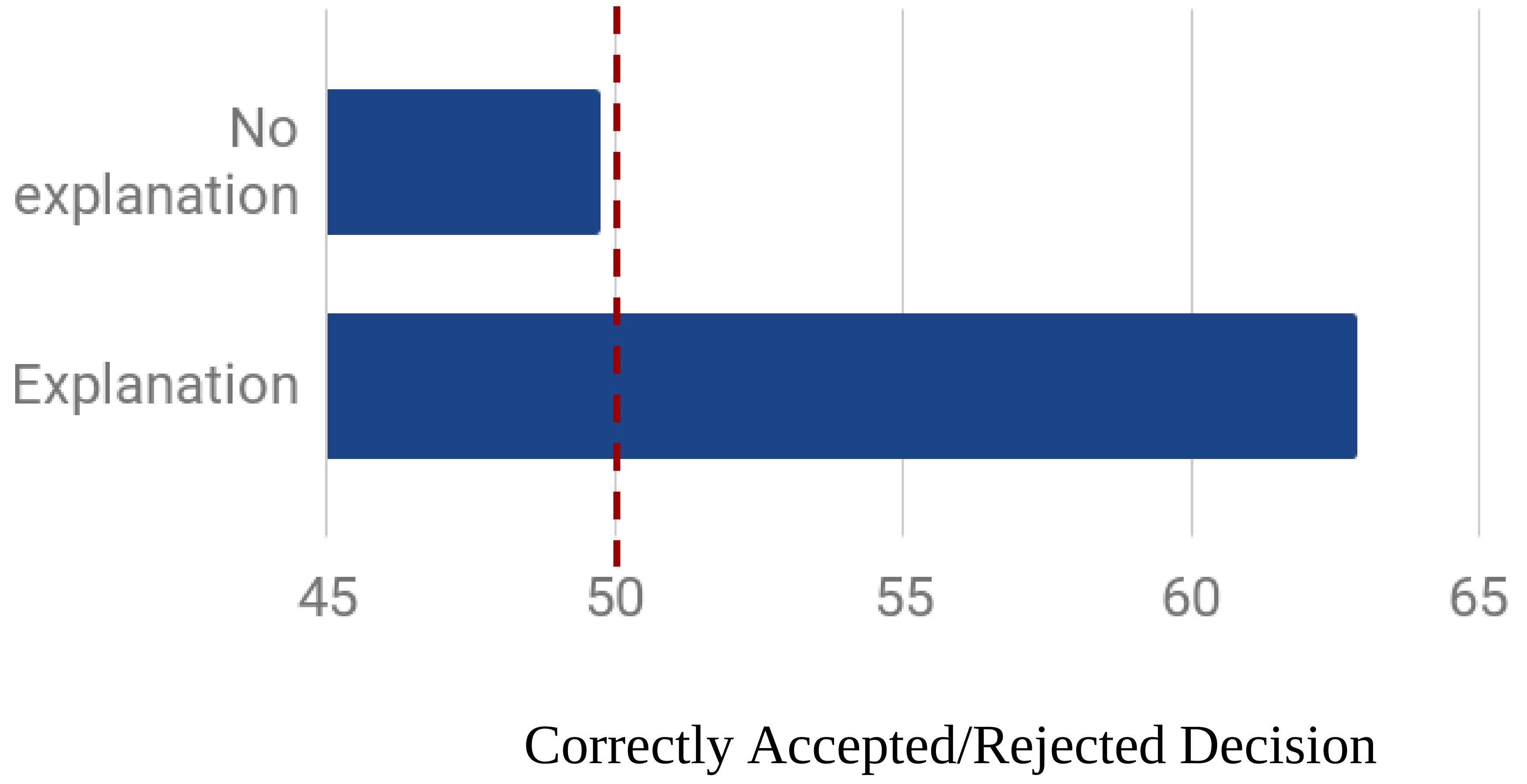
Are Explanations Helpful to Humans?



Are Explanations Helpful to Humans?

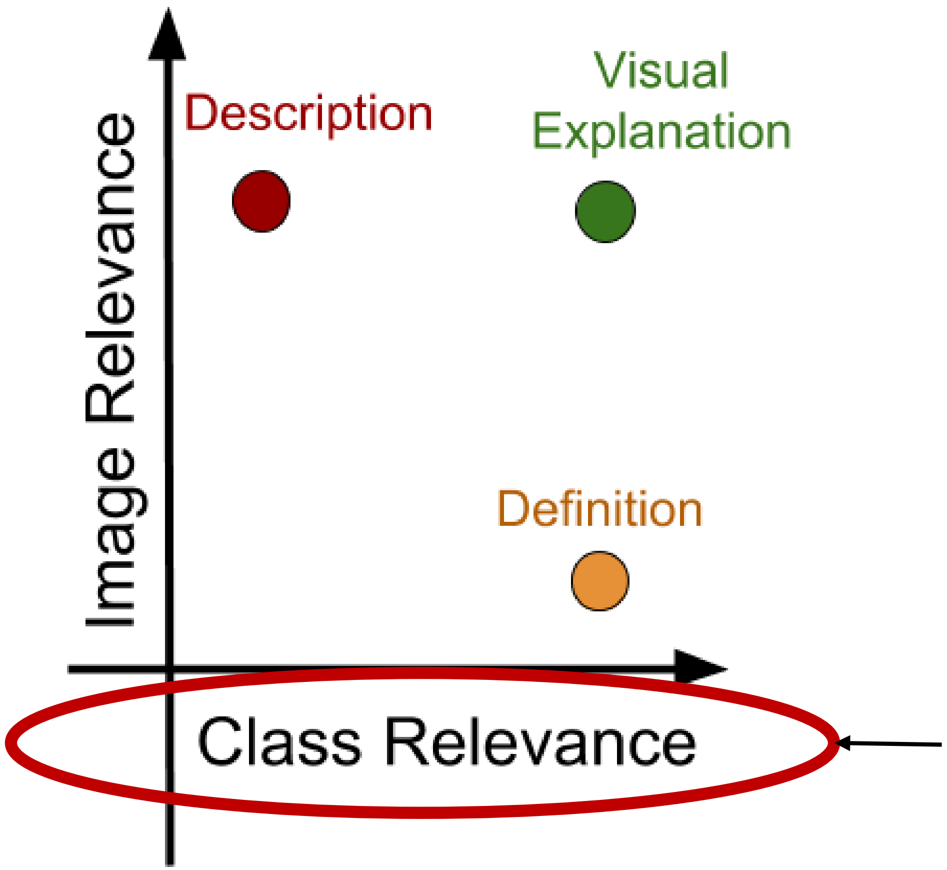


Are Explanations Helpful to Humans?



What makes a good visual explanation?

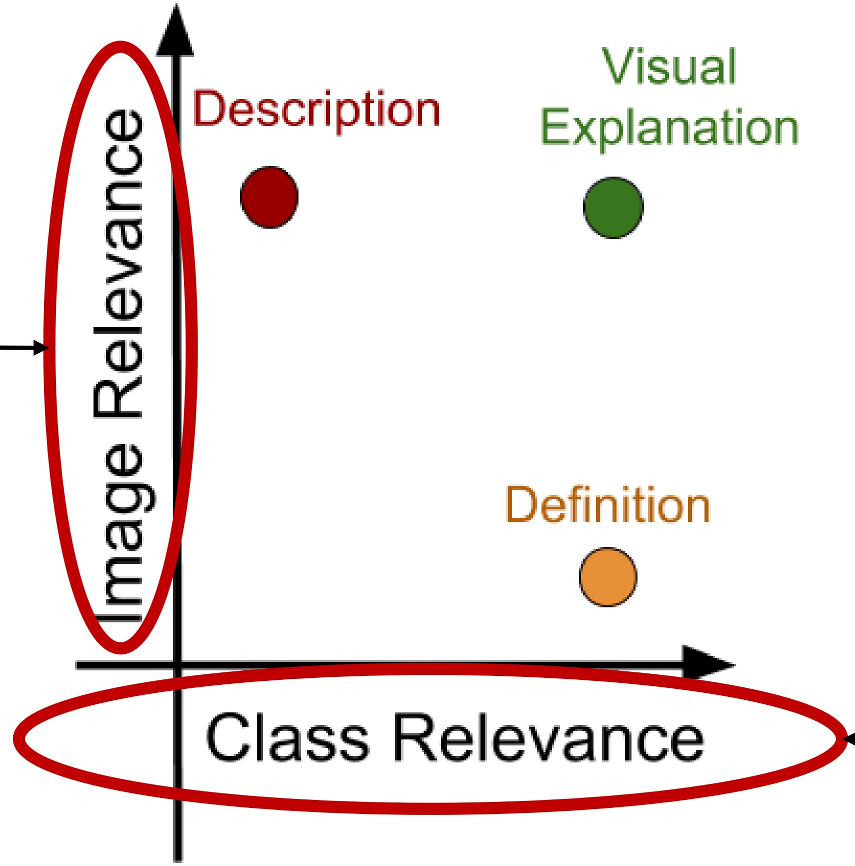
What makes a good visual explanation?



Discriminative Loss

What makes a good visual explanation?

Grounding



Discriminative
Loss

Driving-X

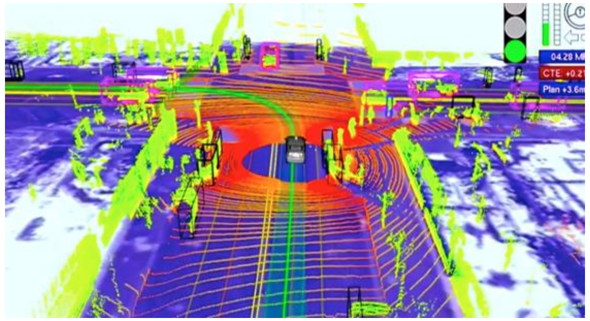


Image credit: Berkeley Deep Drive

Image credit: H. Miller, 1957

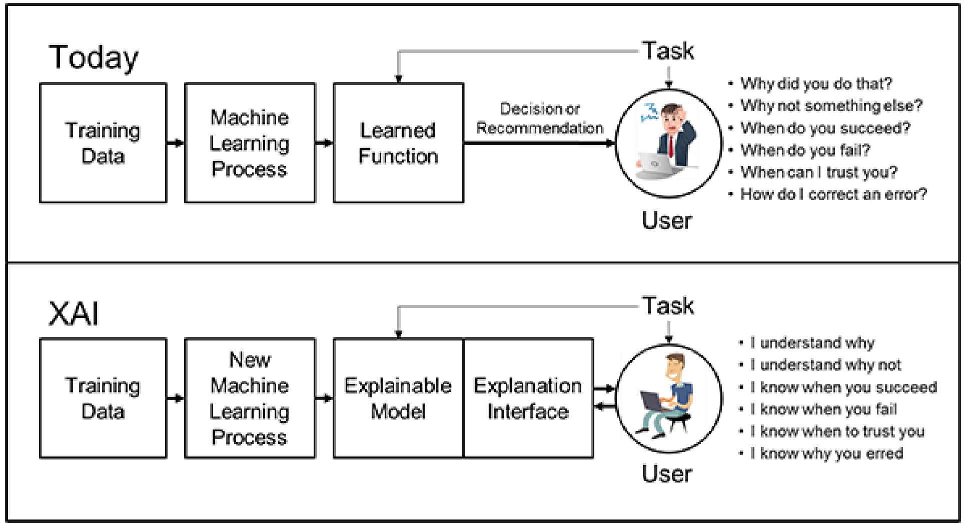
Jinkyu Kim, Anna Rohrbach,
Trevor Darrell, John Canny,
and Zeynep Akata

UC Berkeley
University of Amsterdam

eXplainable AI (for self-driving cars)

Need ***introspective*** or ***debuggable*** driving model:

Explanations are grounded in the network's true internal state.



[DARPA-BAA-16-53, 2017]

Why?

1) Requires a very high level of **trust.**

2) Users should be able to **anticipate** what the vehicle will do.

3) Effective human-machine **communication.**

Outline

Phase 1

- ❑ **Interpretable Learning for Self-driving Cars by Visualizing Causal Attention**

  Jinkyu Kim and John Canny, ICCV 2017.

- ❑ **Textual Explanations for Self-driving Vehicles**

  Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata, ECCV 2018.

Phase 2

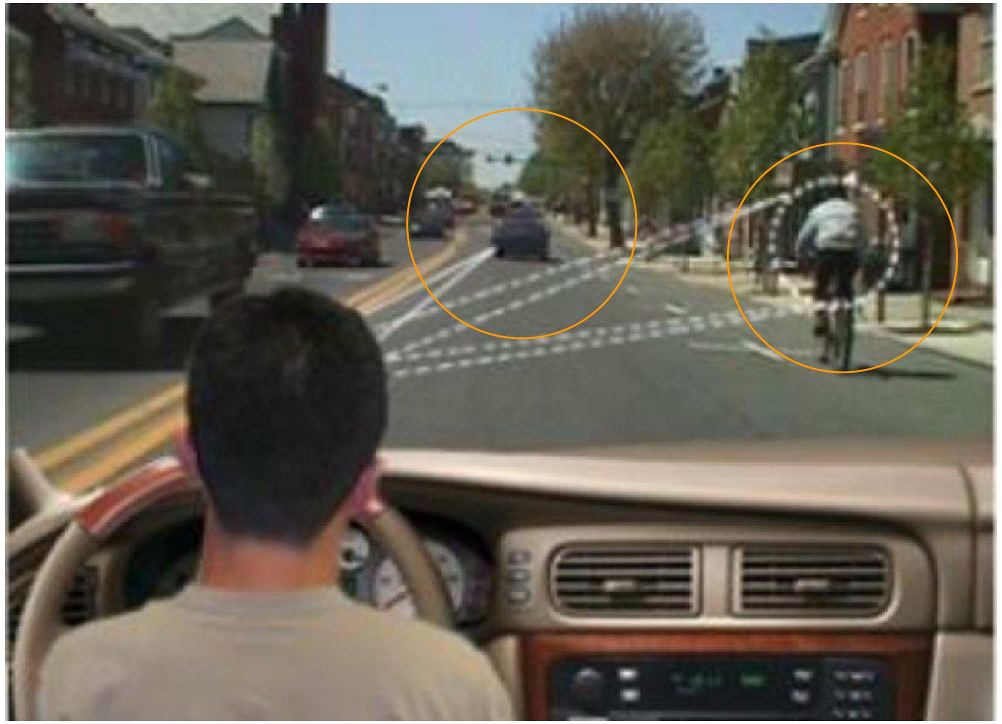
- ❑ **Internalizing Human-to-Vehicle Advice for Self-driving Vehicles**

  Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny, *CVPR 2019*.

- ❑ **Advisable Learning for Self-driving Vehicles**

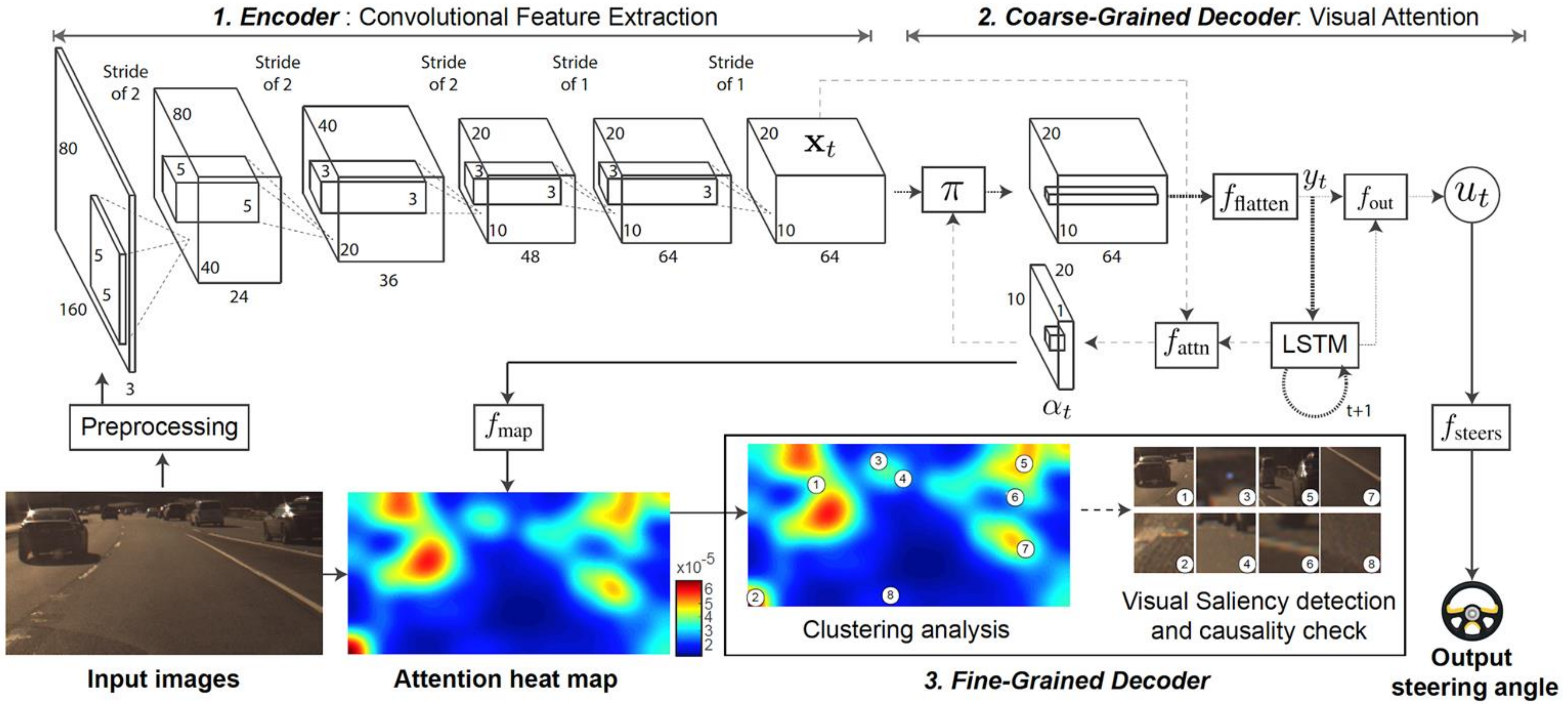
  Jinkyu Kim, Anna Rohrbach, Dequan Wang, Trevor Darrell, and John Canny, *under review*.

Visualizing Causal Attention



Visualizing Causal Attention

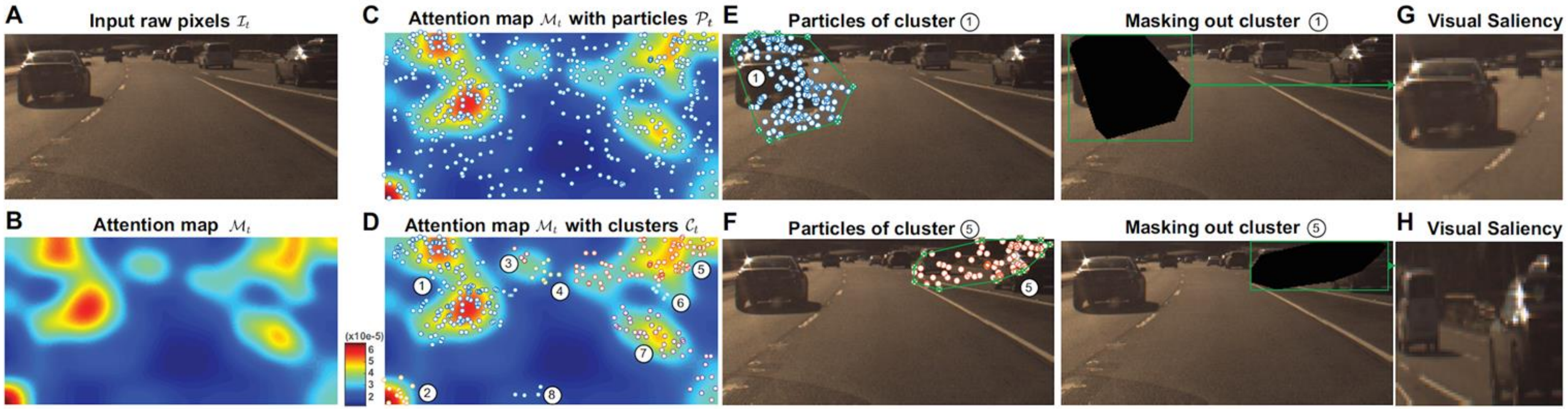
Highlights image regions that causally influence the network's output (i.e., steering)



[Kim and Canny, ICCV'17]

Fine-Grained Decoder (Causality check)

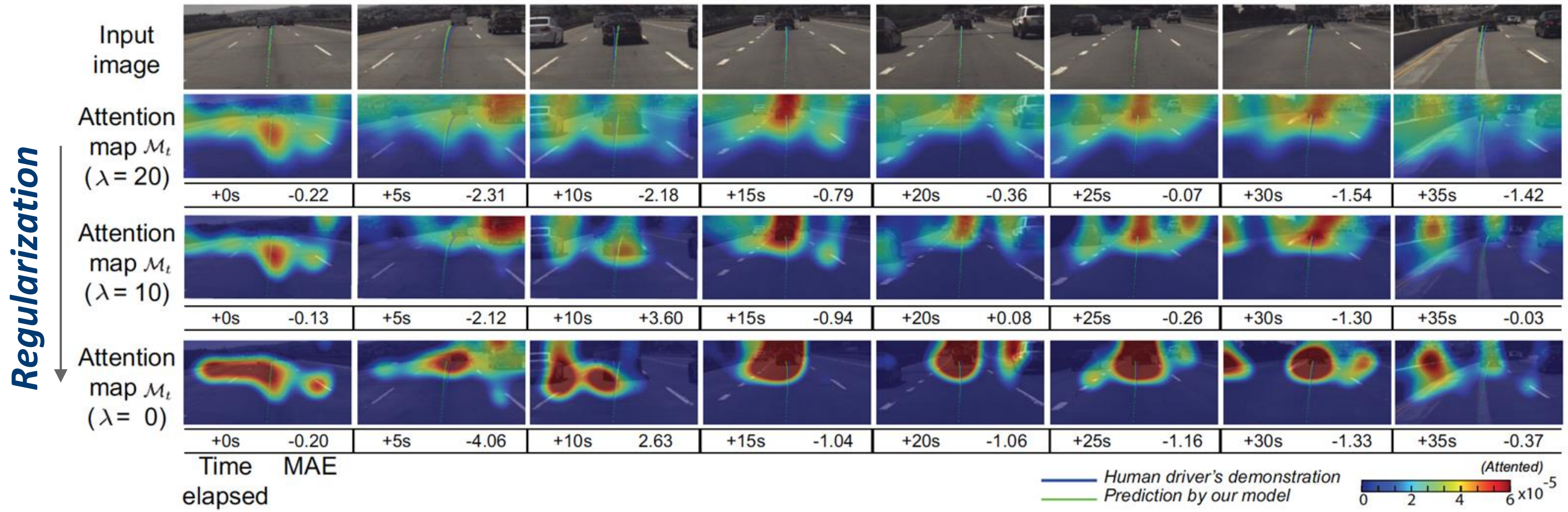
Fine-grained decoder to remove spurious attention blobs and to find causal local visual blobs



[Kim and Canny, ICCV'17]

Examples of Attention Map

Attention maps over time (from left to right)



[Kim and Canny, ICCV'17]

Quantitative Evaluation (Goodness)

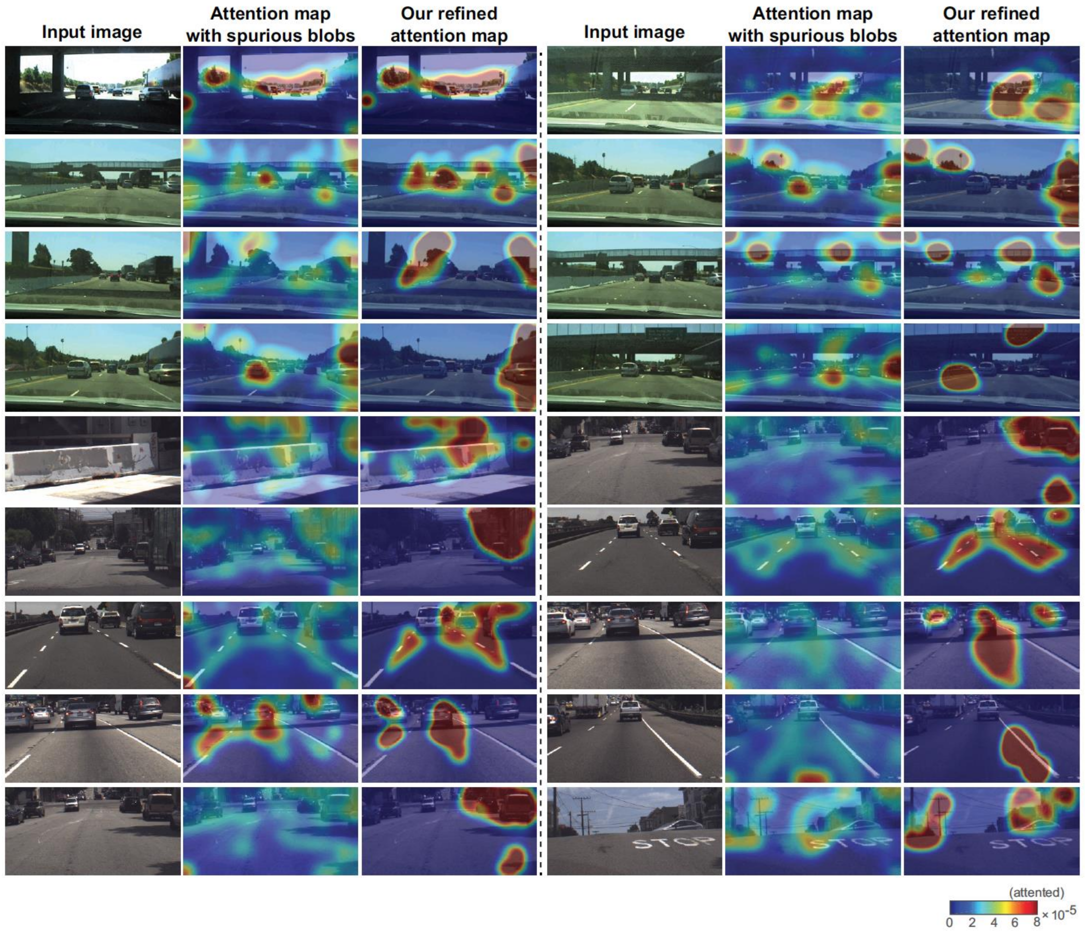
Control accuracy is ***not degraded*** by incorporation of attention compared to an identical base CNN without attention.

Dataset	Model	MAE in degree [SD] Training	MAE in degree [SD] Testing
Comma.ai	CNN+FCN	.421 [0.82]	2.54 [3.19]
	CNN+LSTM	.488 [1.29]	2.58 [3.44]
	Attention (λ=0)	.497 [1.32]	2.52 [3.25]
	Attention (λ=10)	.464 [1.29]	2.56 [3.51]
	Attention (λ=20)	.463 [1.24]	**2.44 [3.20]**
HCE	CNN+FCN	.246 [.400]	1.27 [1.57]
	CNN+LSTM	.568 [.977]	1.57 [2.27]
	Attention (λ=0)	.334 [.766]	**1.18 [1.66]**
	Attention (λ=10)	.358 [.728]	1.25 [1.79]
	Attention (λ=20)	.373 [.724]	1.20 [1.66]
Udacity	CNN+FCN	.457 [.870]	**4.12 [4.83]**
	CNN+LSTM	.481 [1.24]	4.15 [4.93]
	Attention (λ=0)	.491 [1.20]	4.15 [4.93]
	Attention (λ=10)	.489 [1.19]	4.17 [4.96]
	Attention (λ=20)	.489 [1.26]	4.19 [4.93]

[Kim and Canny, ICCV'17]

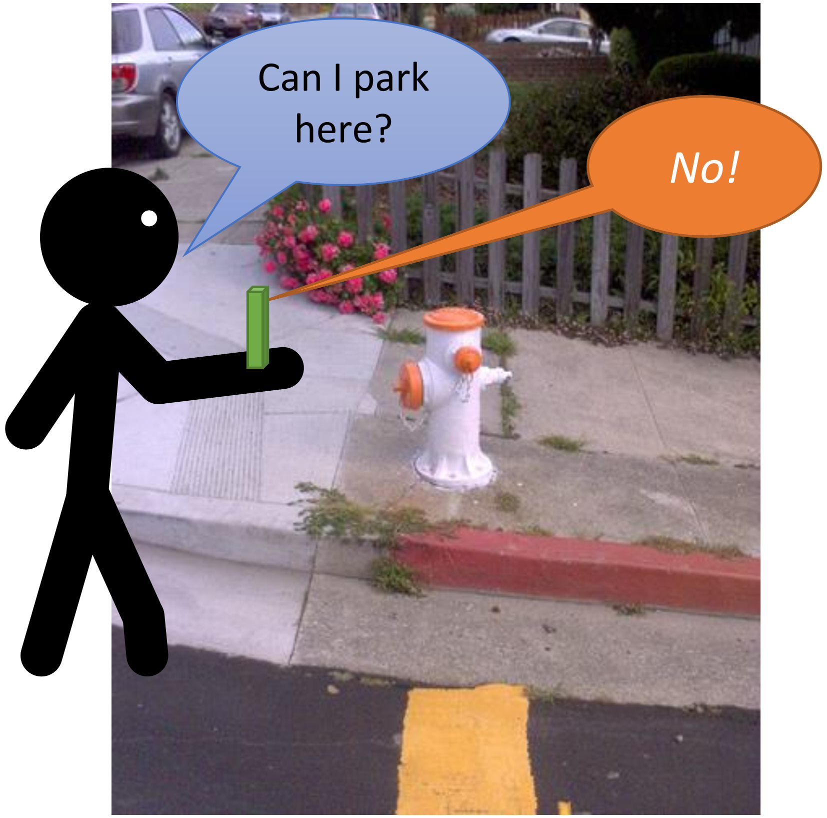
Causal Attention Heat Maps

- ❑ Raw input image
- ❑ Visual attention heatmaps ***with spurious*** attention sources
- ❑ Attention heat maps by ***filtering out spurious blobs***



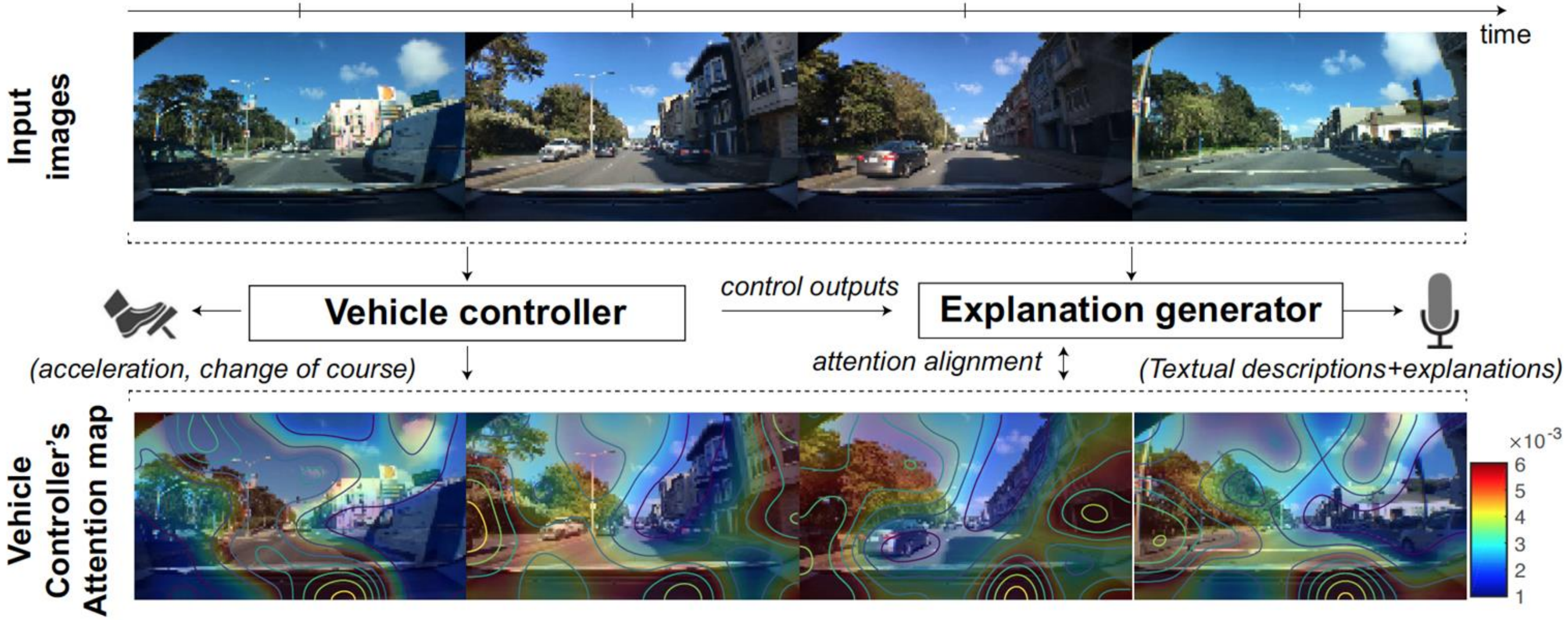
[Kim and Canny, ICCV'17]

Textual Explanations



[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Textual Explanations



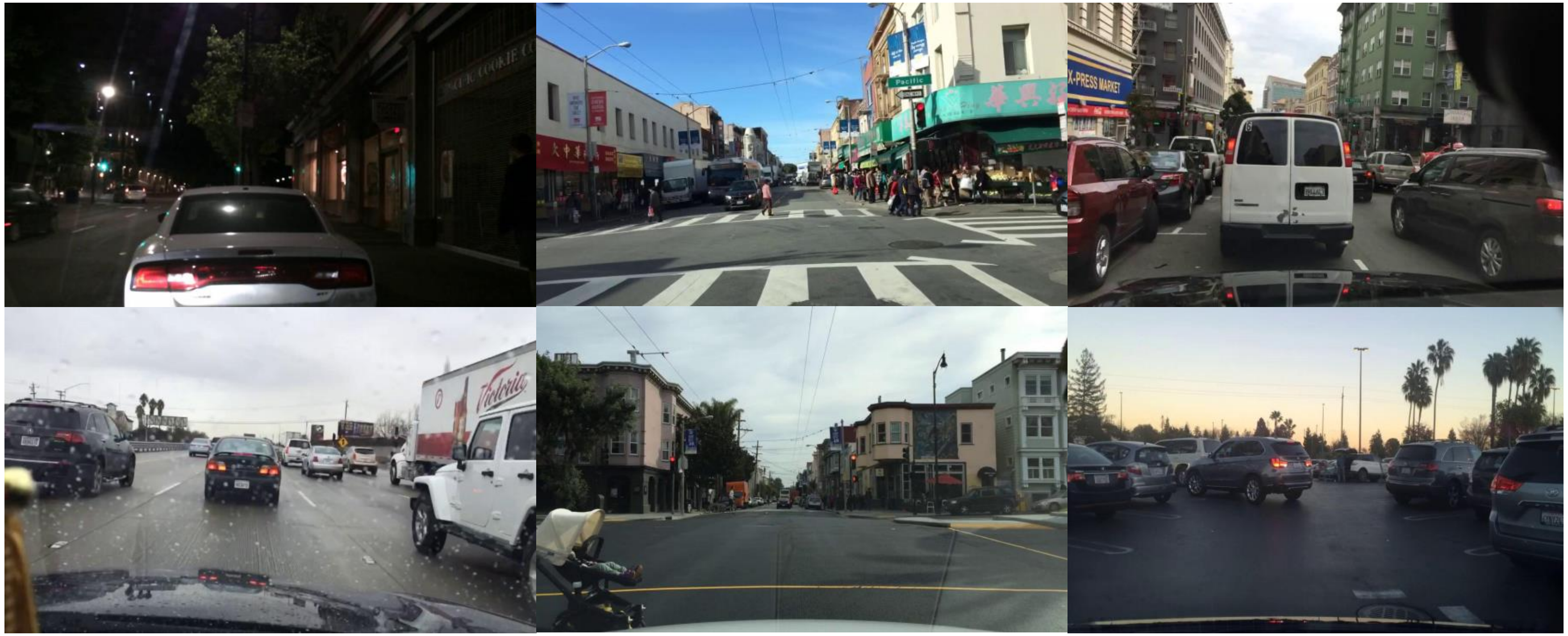
Example of textual descriptions **+** *explanations:*

Ours: *"The car is driving forward* **+** *because there are no other cars in its lane"*

Human annotator: *"The car heads down the street* **+** *because the street is clear."*

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

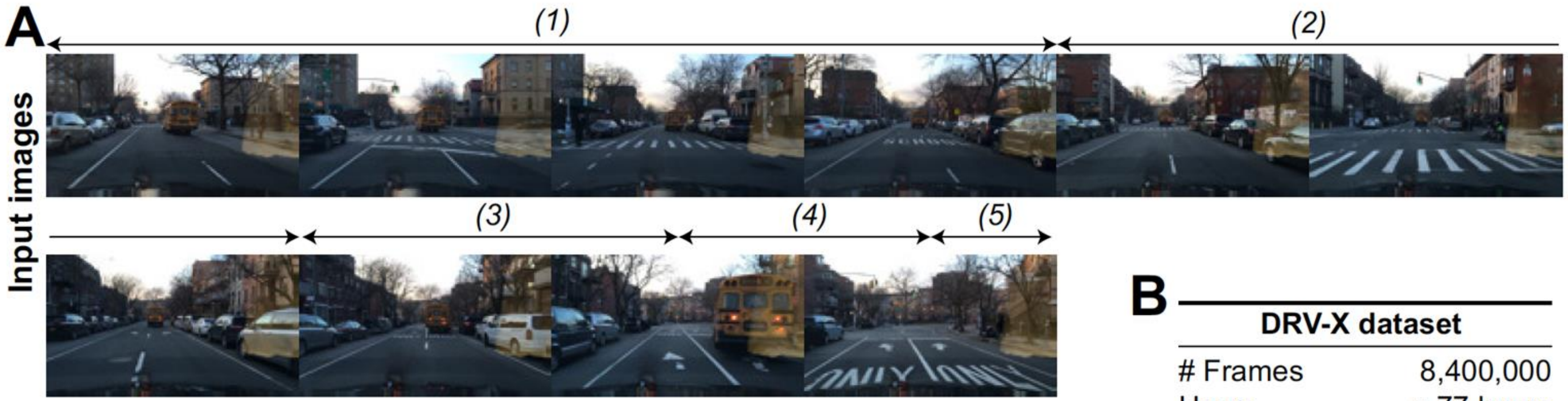
Berkeley DeepDrive Video (BDD-V) Data



Over 10,000 hours of driving data, which provides
(1) dash-cam video, (2) GPS, (3) course and speed

[Xu, Gao, Yu, Darrell, CVPR'17]

Berkeley DeepDrive eXplanation (BDD-X) dataset



Action descriptions:

(1) The car is driving
(2) The car is moving into the right lane
(3) The car moves back into the left lane
(4) The car drives in the left lane
(5) The car moves into the right lane

Action explanations:

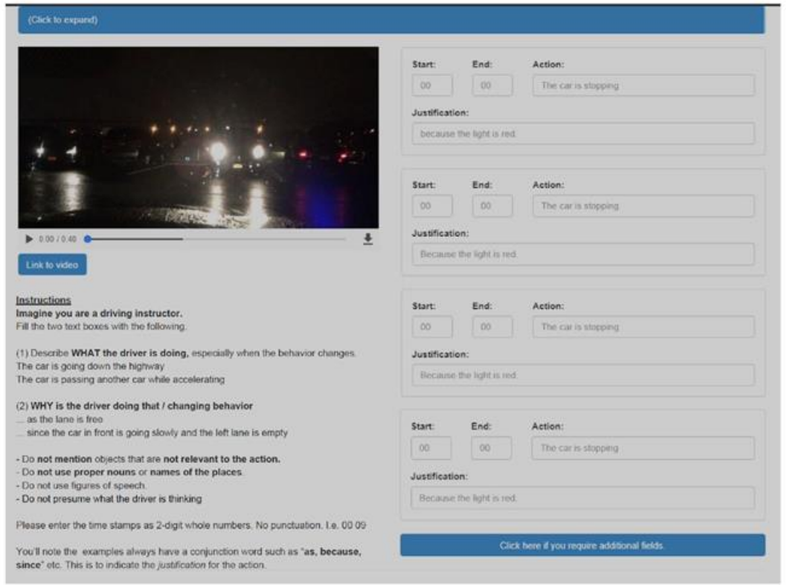
as *there is nothing to impede it.*
because *it is safe to do so.*
because *the school bus in front of it is stopping.*
in order to *pass the school bus.*
since *it has now passed the school bus and it is taking the right fork.*

B

DRV-X dataset	
# Frames	8,400,000
Hours	≈ 77 hours
Condition	Urban
Lighting	Day/Night
# Annotations	26,228
Avg. # actions / videos	3.8
# Videos	6,984
# Training	5,588
# Validation / Testing	698

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Berkeley DeepDrive eXplanation (BDD-X) dataset

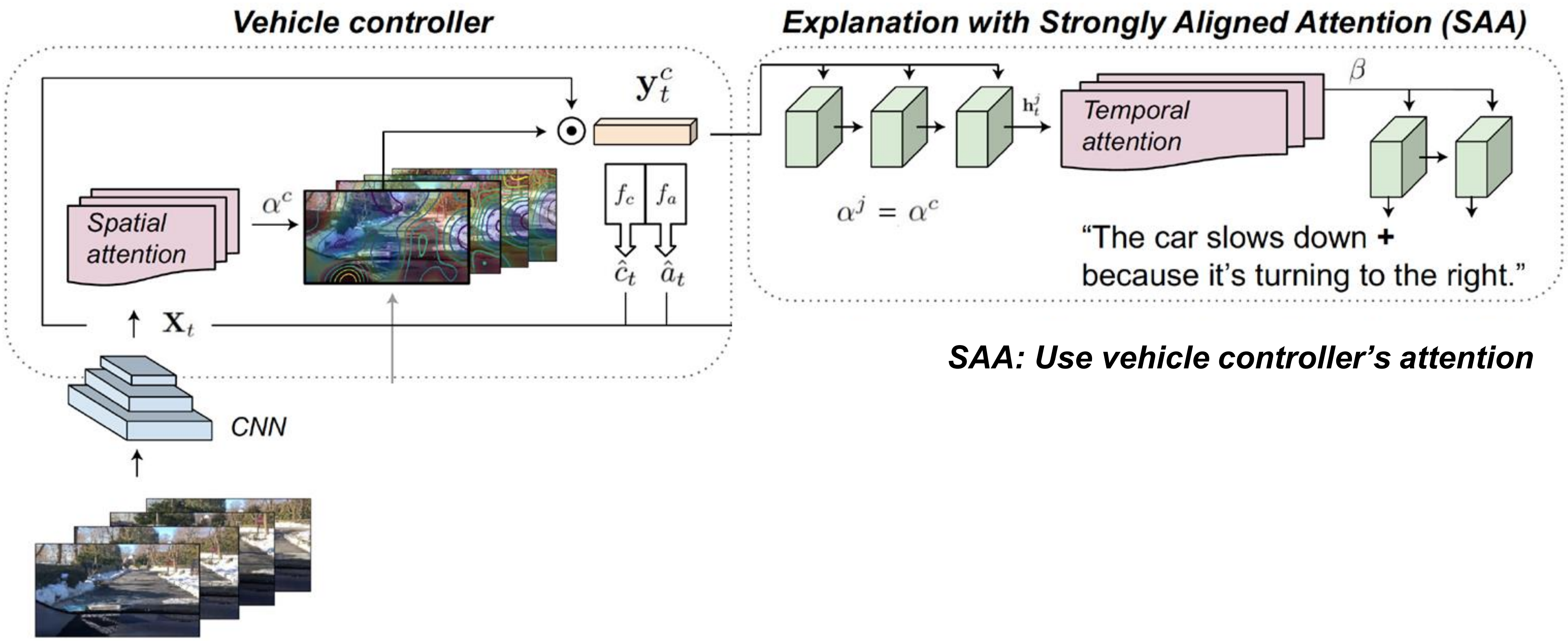


BDD-X action descriptions		BDD-X action explanations	
Word	Count	Word	Count
stop	6879	traffic	7486
slow	6122	light	6116
forward	4322	red	3979
drive	3994	move	3915
move	3273	clear	3660
accelerate	2882	ahead	3629
right	2616	road	3528
left	2574	stop	3430

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Model

Two approaches (SAA and WAA) to align the vehicle controller and the textual justifier such that they look at the same input regions.

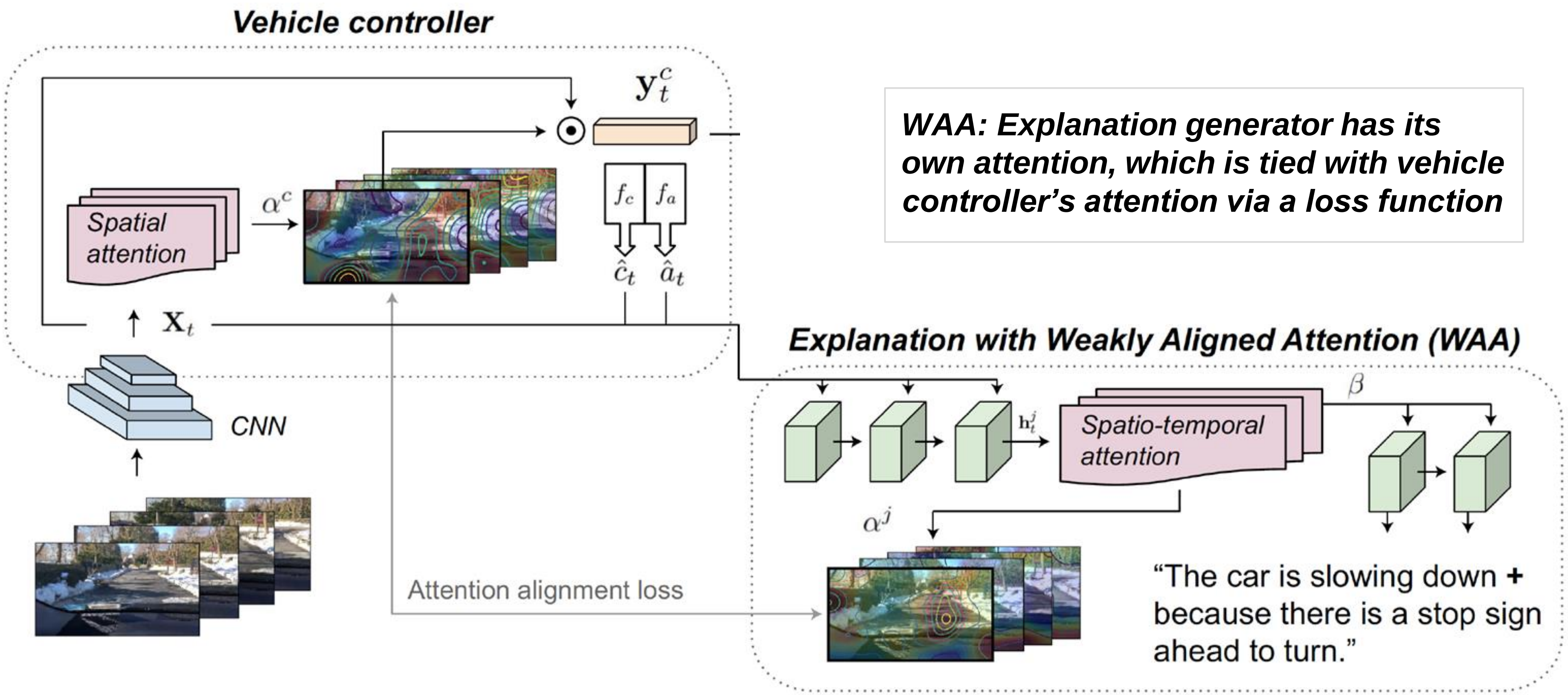


SAA: Use vehicle controller's attention

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Model

Two approaches (SAA and WAA) to align the vehicle controller and the textual justifier such that they look at the same input regions.



WAA: Explanation generator has its own attention, which is tied with vehicle controller's attention via a loss function

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Quantitative Analysis

Model	λ_c	Mean of absolute error (MAE)		Mean of distance correlation	
		Acceleration (m/s^2)	Course (degree)	Acceleration (m/s^2)	Course (degree)
CNN+FC [1]†	-	6.92 [7.50]	12.1 [19.7]	0.17 [0.15]	0.16 [0.14]
CNN+FC [1]+P	-	6.09 [7.73]	6.74 [14.9]	0.21 [0.18]	0.39 [0.33]
CNN+LSTM+Attention [4]†	-	6.87 [7.44]	10.2 [18.4]	0.19 [0.16]	0.22 [0.18]
CNN+LSTM+Attention+P (Ours)	1000	5.02 [6.32]	6.94 [15.4]	0.65 [0.25]	0.43 [0.33]
CNN+LSTM+Attention+P (Ours)	100	2.68 [3.73]	6.17 [14.7]	0.78 [0.28]	0.43 [0.34]
CNN+LSTM+Attention+P (Ours)	10	2.33 [3.38]	6.10 [14.7]	0.81 [0.27]	0.46 [0.35]
CNN+LSTM+Attention+P (Ours)	0	**2.29 [3.33]**	**6.06 [14.7]**	**0.82 [0.26]**	**0.47 [0.35]**

- **Prior measurements (P) help**
- **Spatial attention helps**
- **Low entropy attention leads to higher error**

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Quantitative Analysis

Type	Model	Control inputs	λ_a	λ_c	Explanations (*e.g.*"*because the light is red*")			Descriptions (*e.g.*"*the car stops*")		
					BLEU-4	METEOR	CIDEr-D	BLEU-4	METEOR	CIDEr-D
	Non-XAI baseline	-	-	-	1.692	8.30	13.29	0.41	21.99	14.17
	S2VT [17]	N	-	-	6.332	11.19	53.35	30.21	27.53	179.8
	S2VT [17]+SA	N	-	-	5.668	10.96	51.37	28.94	26.91	171.3
	S2VT [17]+SA+TA	N	-	-	5.847	10.91	52.74	27.11	26.41	157.0
Rationalization	Ours (no constraints)	Y	0	0	6.515	12.04	61.99	31.01	28.64	205.0
Introspective explanation	Ours (with SAA)	Y	-	0	6.998	12.08	62.24	**32.44**	29.13	213.6
	Ours (with SAA)	Y	-	10	6.760	12.23	63.36	29.99	28.26	203.6
	Ours (with SAA)	Y	-	100	7.074	12.23	66.09	31.84	29.11	214.8
	Ours (with WAA)	Y	10	0	6.967	12.14	64.19	32.24	29.00	**219.7**
	Ours (with WAA)	Y	10	10	6.951	**12.34**	68.56	30.40	28.57	206.6
	Ours (with WAA)	Y	10	100	**7.281**	12.24	**69.52**	32.34	**29.22**	215.8

- **All better than baseline**
- **Introspective better than Rationalization**
- **WAA is best**

Abbreviation:
S2VT (seq-to-seq video-to-text)
TA (temporal fusion)
SA (spatial attention)
WAA (Weakly aligned attention)
SAA (Strongly aligned attention)

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

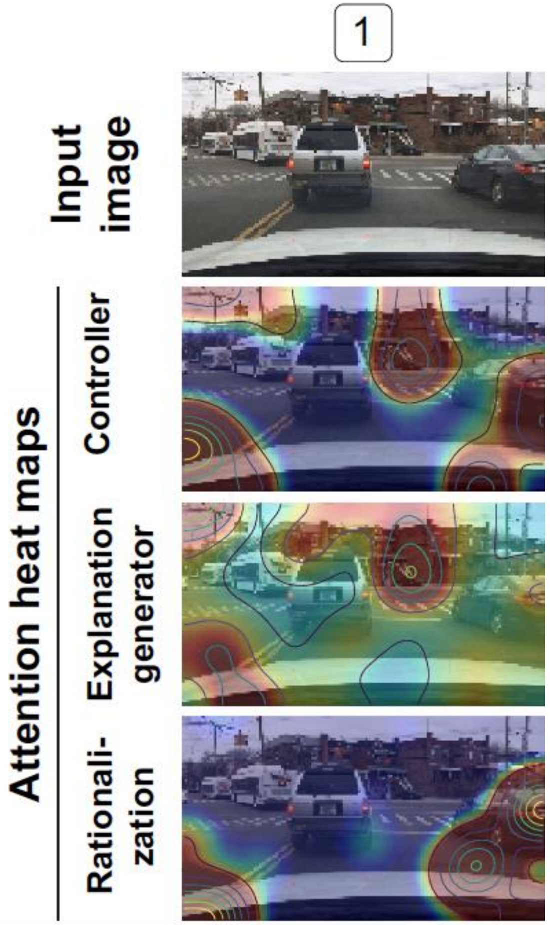
Human Evaluation

Type	Model	Control inputs	λ_a	λ_c	Correctness rate	
					Explanations	Descriptions
Non-XAI baseline†		-	-	-	22.4%	35.6%
Rationalization	Ours (no constraints)	Y	0	0	64.0%	92.8%
Introspective explanation	Ours (with SAA)	Y	-	100	62.4%	90.8%
	Ours (with WAA)	Y	10	100	**66.0%**	**93.5%**

Table 3: Human evaluation of the generated action descriptions and explanations for randomly chosen 250 video intervals. We measure the success rate where at least 2 human judges rate the generated description or explanation with a score 1 (correct and specific/detailed) or 2 (correct). †: Sentences are sampled based on their frequency in the training data (i.e. a strong prior).

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Examples of explanations generated

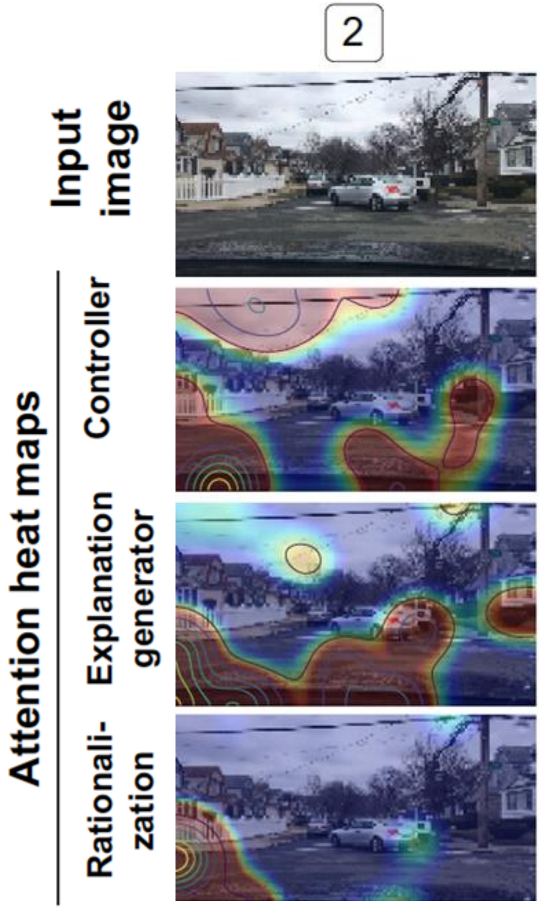


1

Human: The car steadily driving **+** now that the cars are moving.
Ours (WAA): The car is driving forward **+** because traffic is moving freely.
Ours (SAA): The car heads down the road **+** because traffic is moving at a steady pace.
Rationalization: The car slows down **+** because it's getting ready to a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Examples of explanations generated

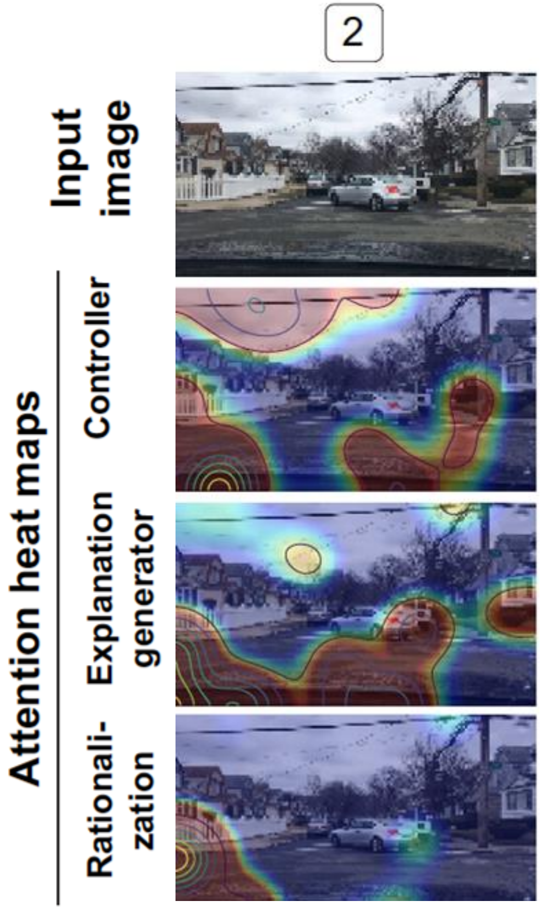


2

Human: The car slows down + since it is about to turn left.
Ours (WAA): The car slows down + because it is preparing to turn to the road.
Ours (SAA): The car is slowing + because it is approaching a stop sign.
Rationalization: The car slows + because there is a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Examples of explanations generated

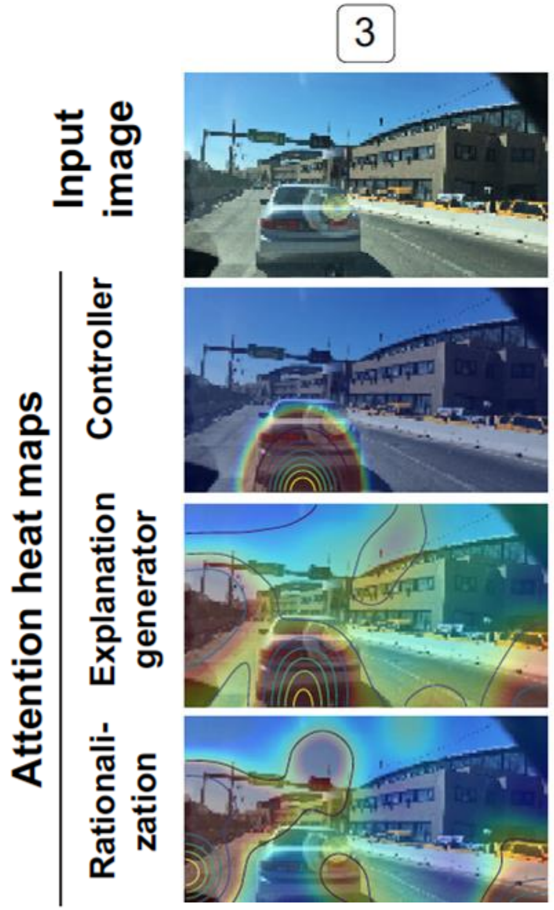


2

Human: The car slows down + since it is about to turn left.
Ours (WAA): The car slows down + because it is preparing to turn to the road.
Ours (SAA): The car is slowing + because it is approaching a stop sign.
Rationalization: The car slows + because there is a stop sign.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

Examples of explanations generated



3

Human: The car is stopped **+** while it waits for traffic in front of it to move.

Ours (WAA): The car is stopped **+** because traffic is stopped.
Ours (SAA): The car is stopped **+** because the car in front of it is stopped.

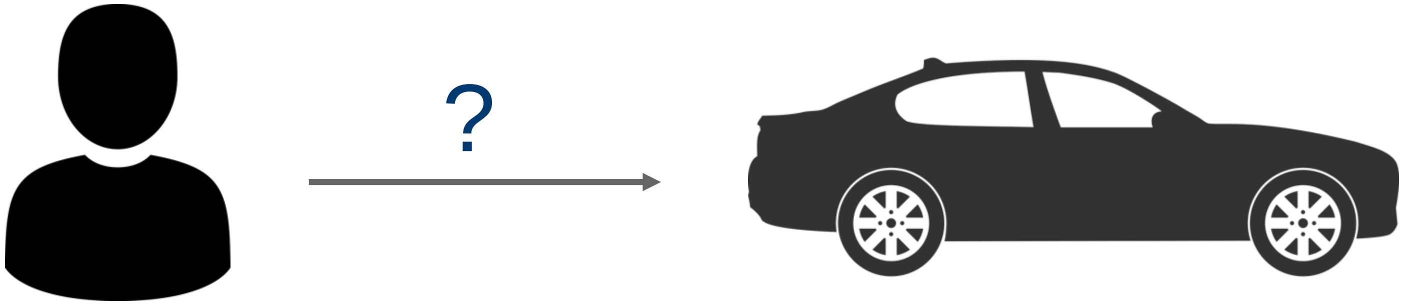
Rationalization: The car is stopped **+** because it is parked in the left lane.

[Kim, Rohrbach, Darrell, Canny, Akata, ECCV'18]

eXplainable AI (for self-driving cars)

- ❑ Visualizing **Attention** Maps
- ❑ Generating Textual **Explanations**

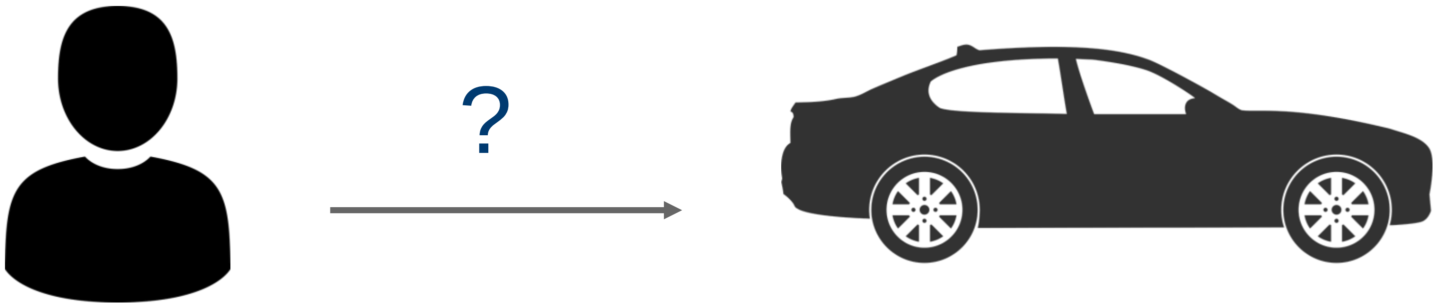
Advisable AI (for self-driving cars)



We want to allow end-users to not only ***understand*** the controller, but to ***influence*** it.

Why Advice and Not Commands?

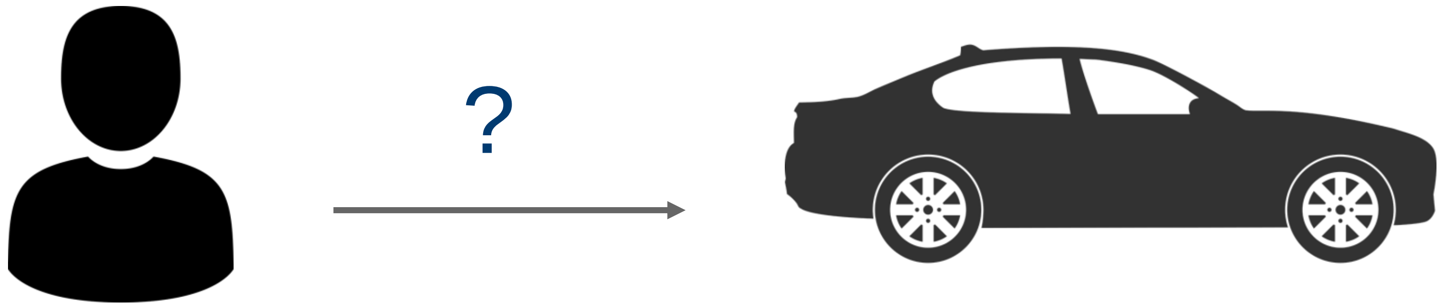
- Users will not be aware of the full state of the vehicle (they are not driving), controller should be in charge.
- Users have real world knowledge that the controller lacks which can improve safety and ride quality.



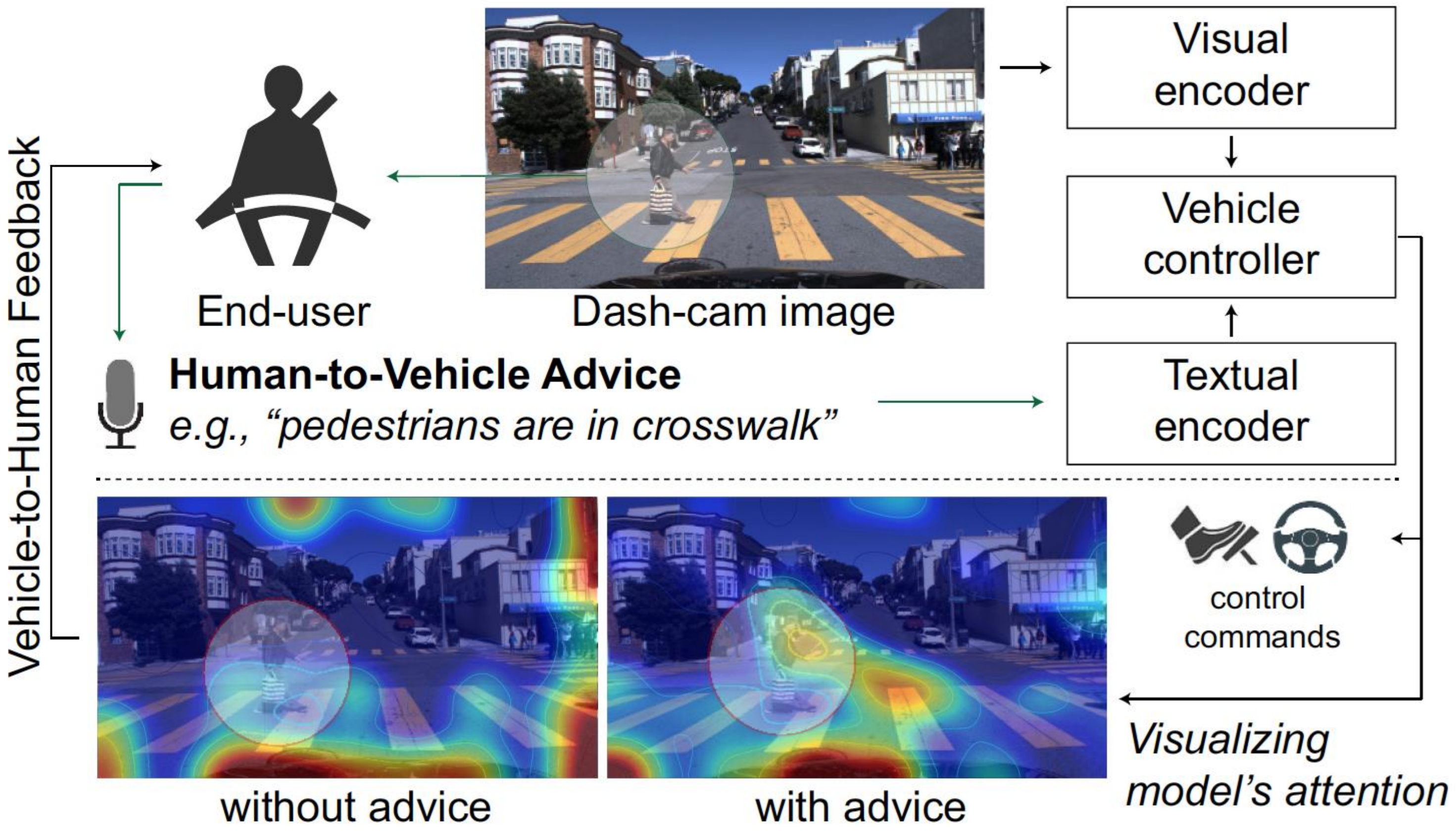
Why Advice and Not Commands?

- Advice will often be given offline, or at the beginning of a ride:
  - When at an intersection, look out for pedestrians
  - Drive gently (occupant gets carsick)

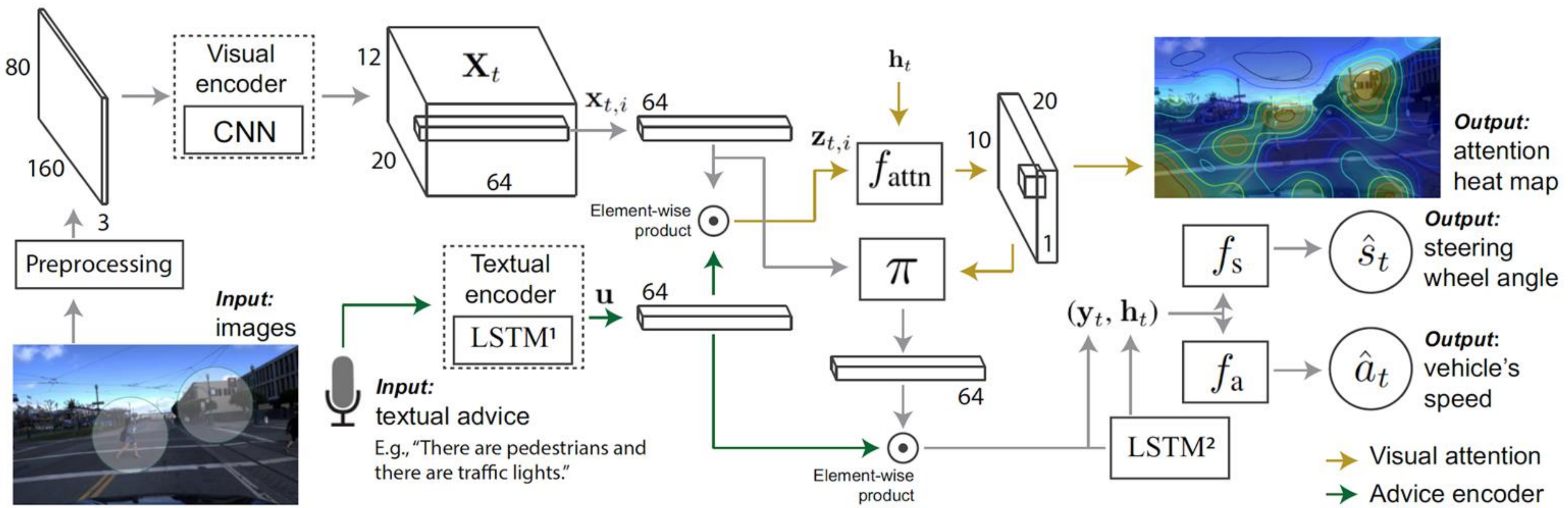
- What we have now: instantaneous control using text grounded in the video.



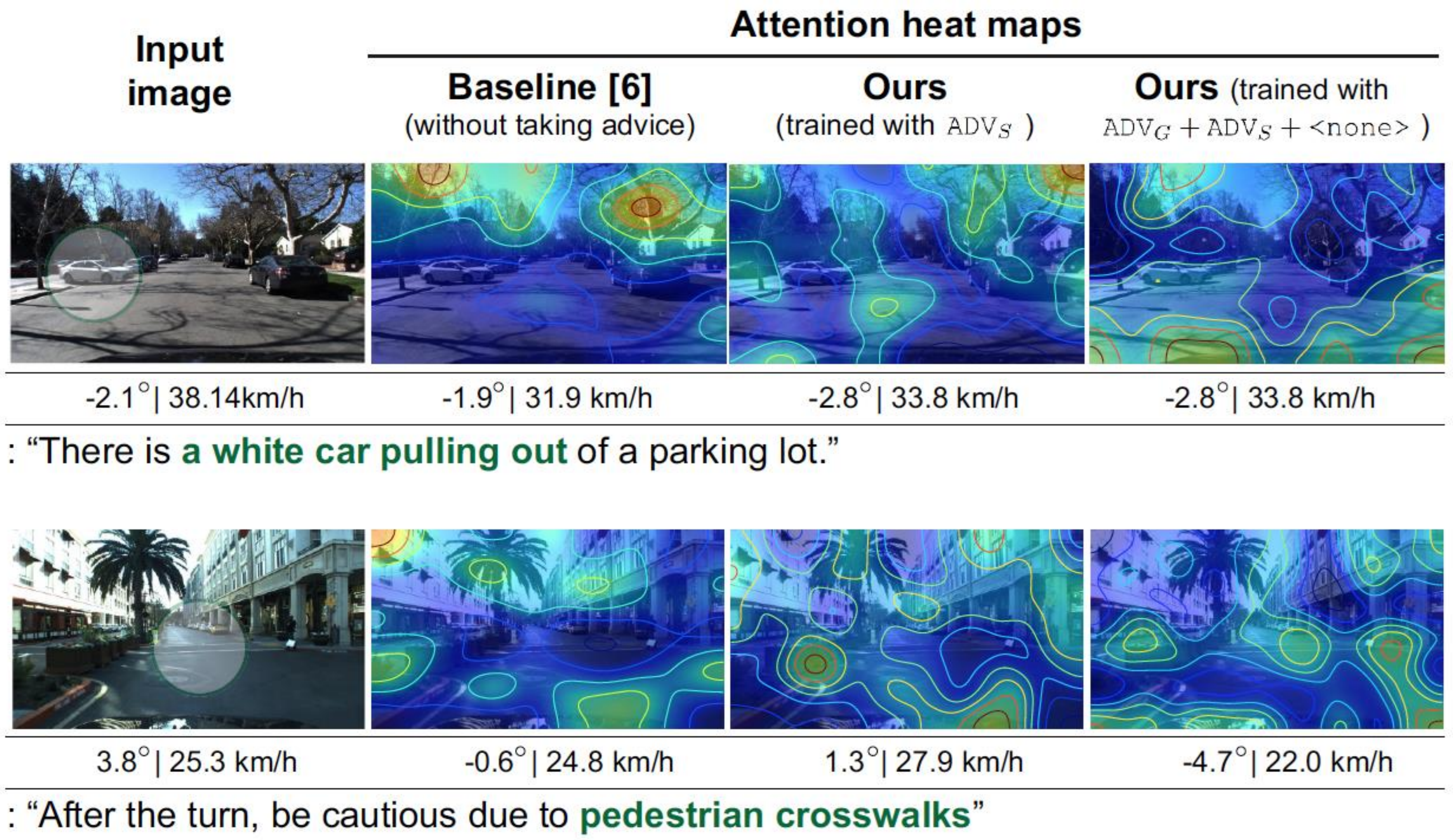
Advisable AI (for self-driving cars)



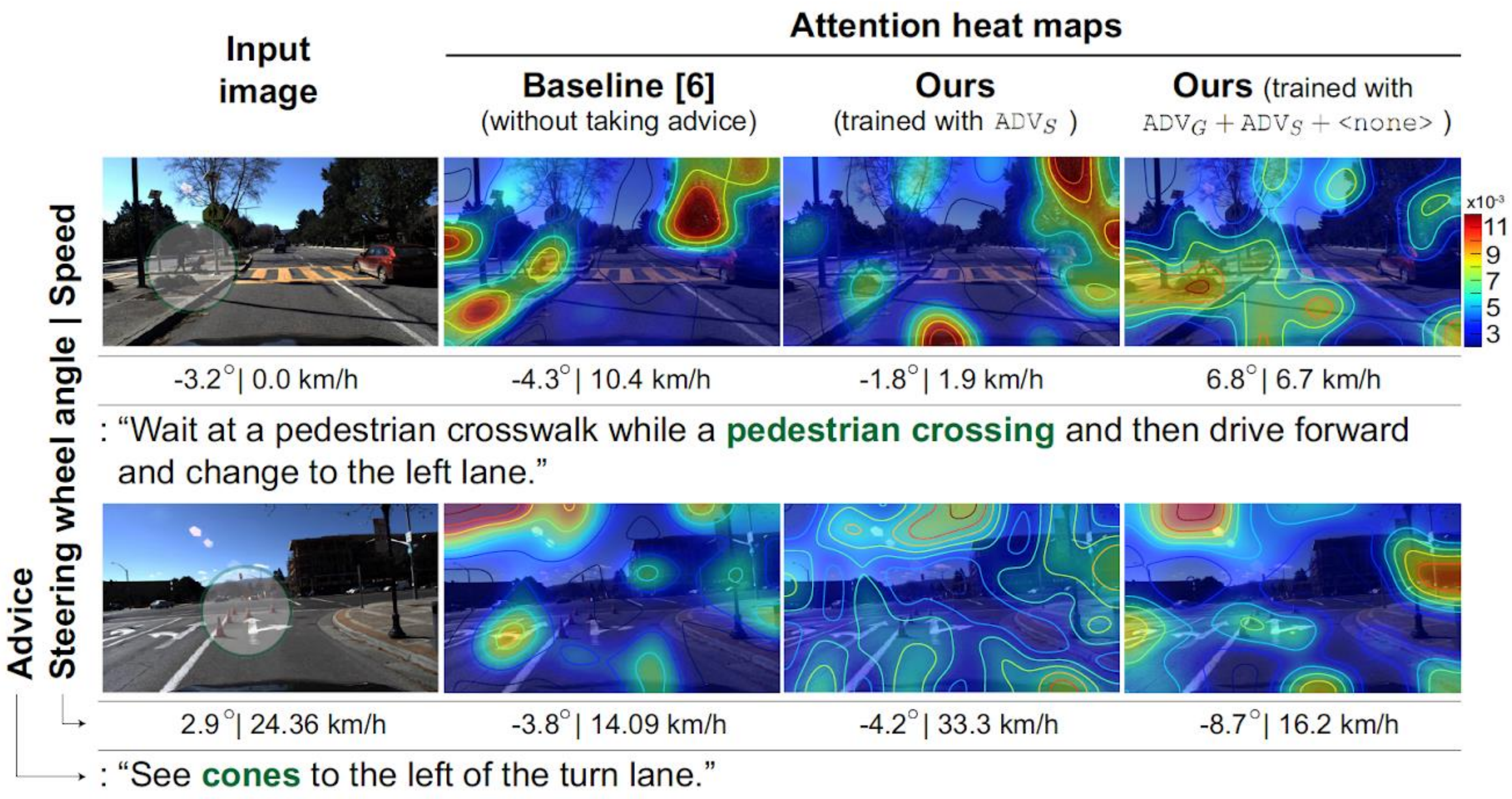
Advisable AI (for self-driving cars)



Qualitative results



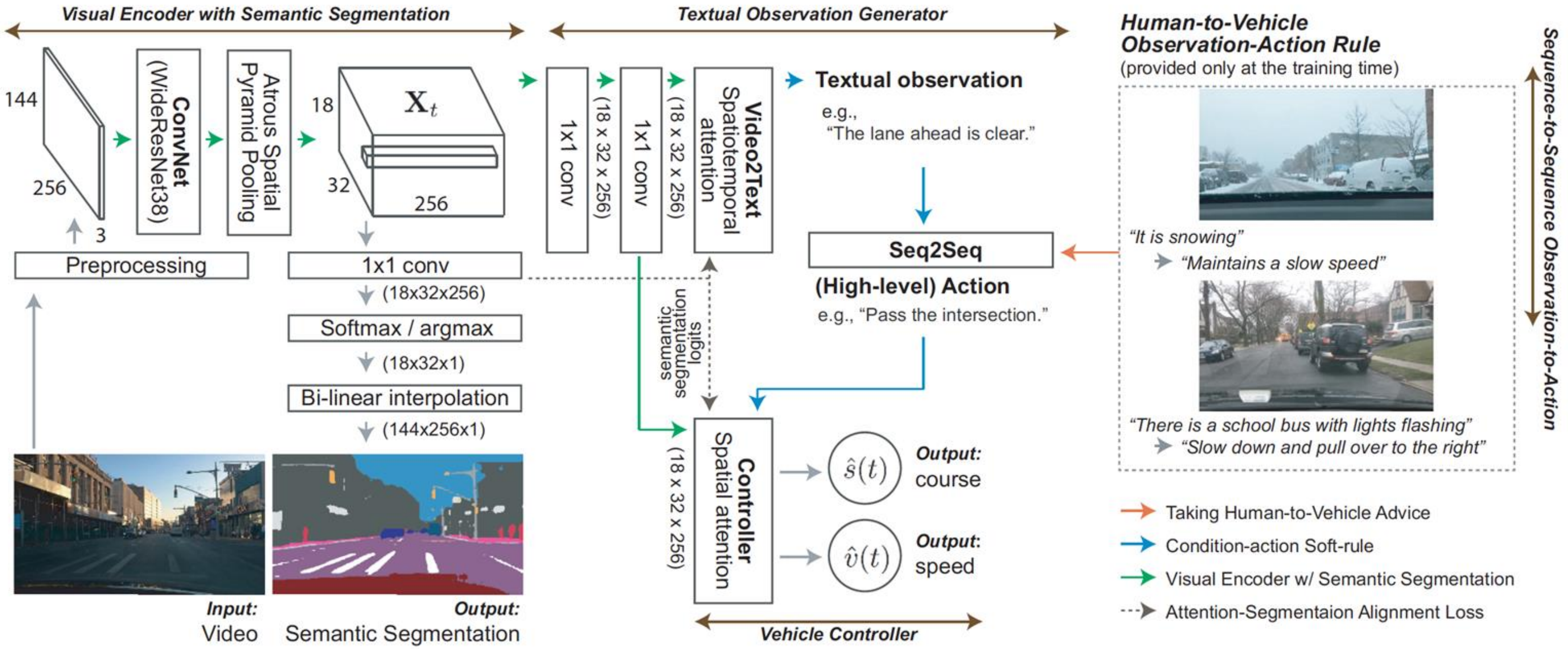
Qualitative results



Fine-grained Attention / Long-term Advice

We have explored an explainable and advisable driving model, which we explore several approaches for better forms of explainability:

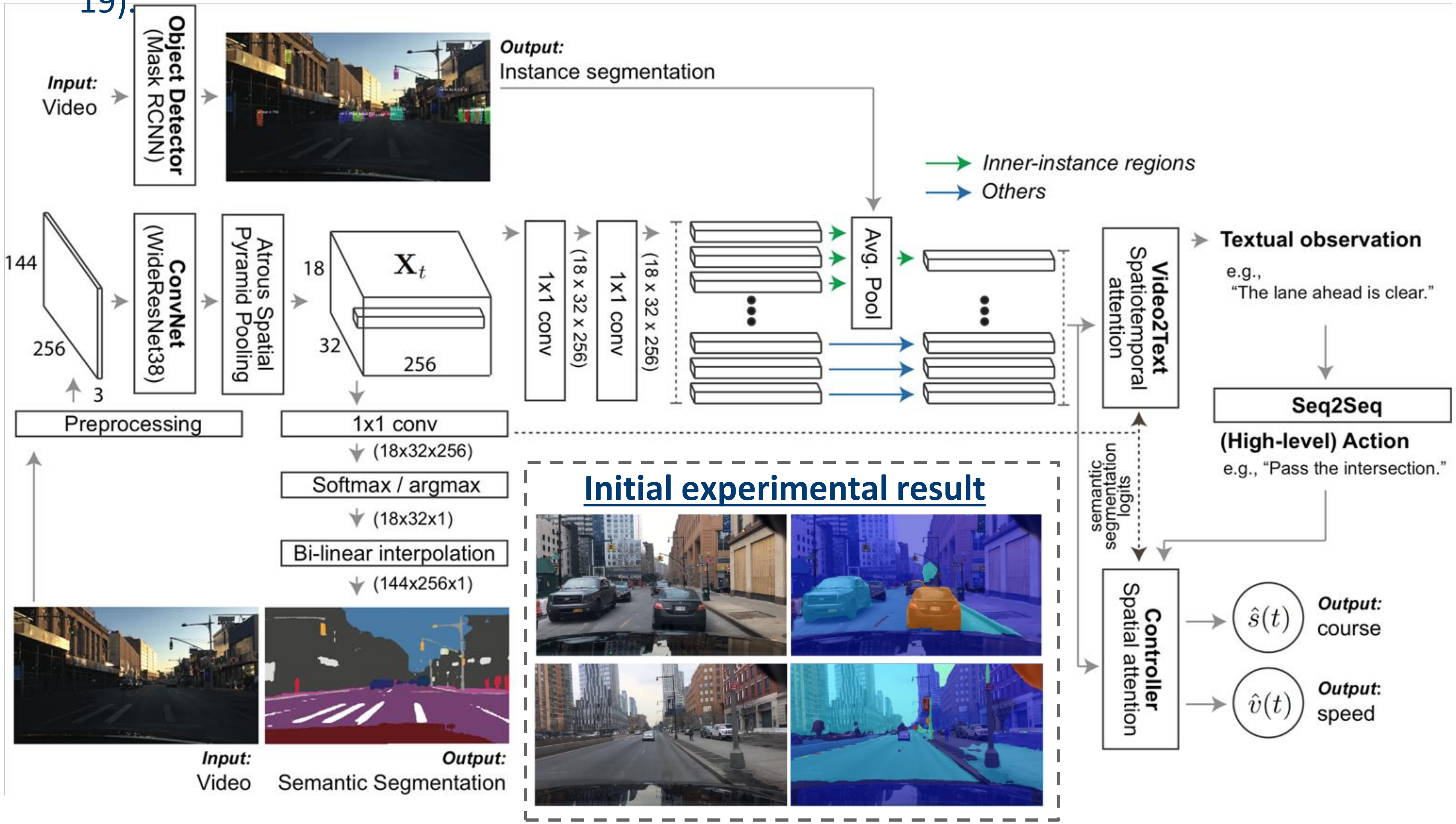
- ❑ We use semantic segmentation as an input representation.
- ❑ We update model to take long-term advice (*i.e* offline driving instructions).



Kim, Rohrbach, Wang, Darrell, and Canny, "Advisable Learning for Self-driving Vehicles," *under review.*

Instance Attention

We are exploring a composite end-to-end vehicle controller that integrates our advisable/explainable model with our recent object-centric attention model (ICRA '19).



Today

- Multi-step Saliency via Compositional NMNs
- Fine-grained Textual Explanations
- From Explainable to “Advisable”Driving Models