

Visual Dialog

Danna Gurari

University of Colorado Boulder
Fall 2022



<https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html>

Review

- Last lecture:
 - Image captioning applications
 - Image captioning datasets
 - Image captioning evaluation
 - Challenge winners
- Assignments (Canvas)
 - Lab assignment 4 due Wednesday
 - Final project proposal due in 2 weeks
- Questions?

Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial


Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial

VQA Dialog

“hold a meaningful dialog with humans in natural language about visual content”

Visual Dialog



A cat drinking water out of a coffee mug.

What color is the mug?

Are there any pictures on it?

Is the mug and cat on a table?

Are there other items on the table?

White and red

No, something is there can't tell what it is

Yes, they are

Yes, magazines, books, toaster and basket, and a plate

Start typing question here ...

VQA Dialog vs VQA and Image Descriptions



VQA

Q: How many people
on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a
wheelchair and one is
holding a racket.

Visual Dialog

Q: How many people are on
wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a
racket ?

A: The woman



Visual Dialog

Q: What is the gender of the
one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

Visual dialog involves memory to answer follow-up questions

Application: Visual Assistance for People with Vision Loss



Applications: Medical

80 year old man s/p vats R lower lobectomy



Q: Airspace opacity?

A: Yes

Q: Fracture?

A: Not in report

Q: Lung lesion?

A: No

Pneumonia?

Yes

Application: Surveillance



Attribute-based Query:

Q: Is it a person in the green bounding box?

A: Yes

(Define the person as P1)

Q: Is P1 female?

A: Yes

Q: Does P1 hold a bag?

A: Yes

Q: Does P1 has long hair and wear leather shoes?

A: Yes

Q: Is P1 in padded jacket and skirt?

A: No

Q: ...

A: ...

Relationship-based Query:

Q: Are they persons in both of the two red bounding boxes?

A: Yes

(Define the upper one as P2, and define the lower one as P3)

Q: Are P2 and P3 the same person?

A: Yes

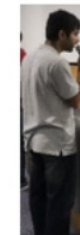
Application: Robotics

e.g., for a companion, psychologist,
and/or assistant in search and
rescue missions (e.g., fire fighters)

User: What are the people doing ?
Nadine: **Talking to each other.**
User: What is their gender ?
Nadine: **Both are male.**



Social Robot: Nadine



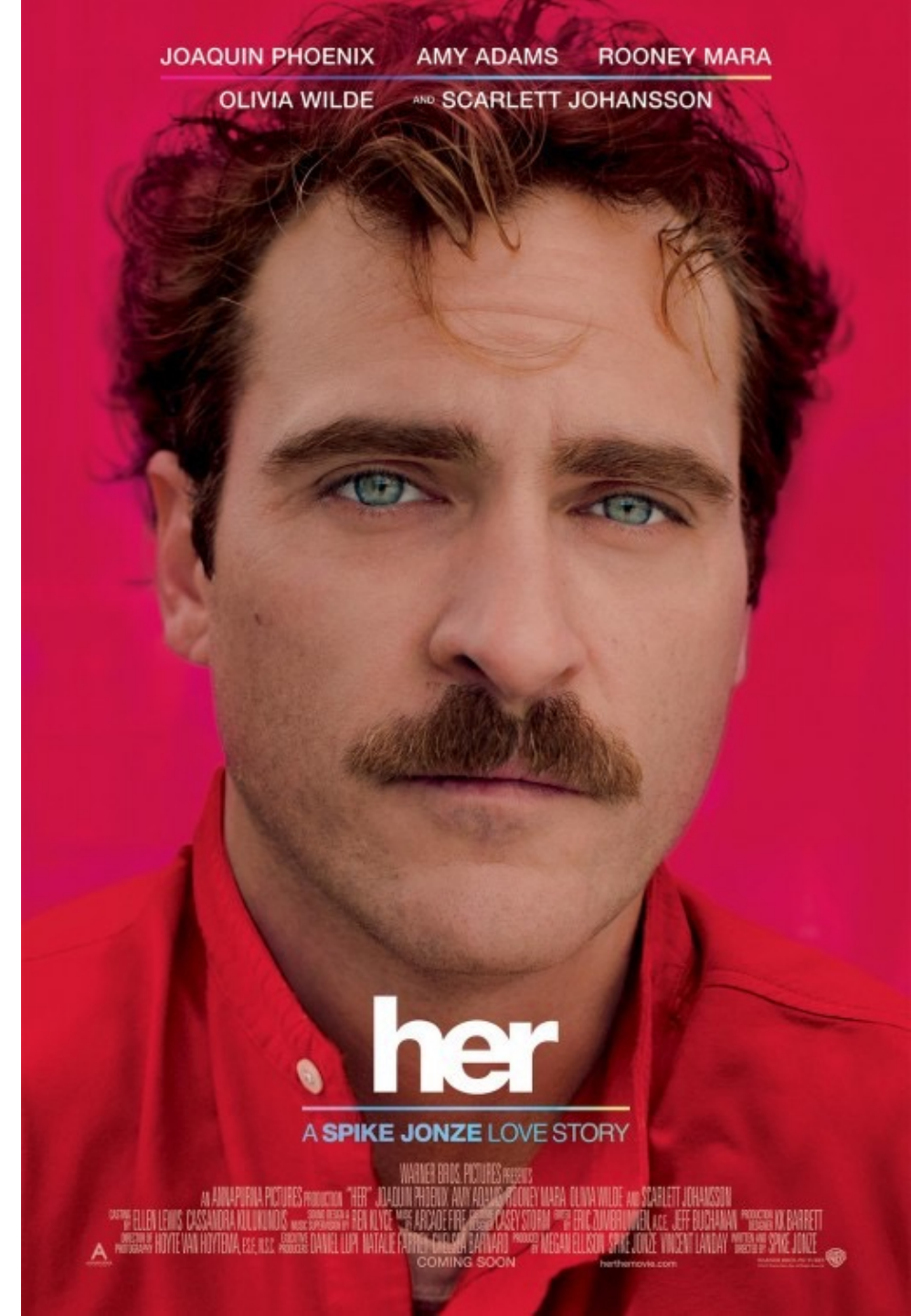
User



Current Scene

Application: Robotics

e.g., for a companion



For what other applications might
visual dialog systems be useful?

Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial

Dataset: Spectrum of Possible Tasks



Popular Datasets

- GuessWhat?!
- VisDial

Popular Datasets

- GuessWhat?!
- VisDial

GuessWhat!? – Crowdsourcing Task

Candidate images restricted to those containing 3-20 objects “to avoid trivial or overly complicated images” and those larger than 500x500 pixels

Target answer
from questioner



2 roles filled by 2 people

Questioner asks yes/no/NA questions until ready to select an answer from a list of options

Questioner

Oracle

Is it a vase?

Yes

Is it partially visible?

No

Is it in the left corner?

No

Is it the turquoise and purple one?

Yes

GuessWhat!? Statistics

- 66,537 images (from COCO)
- 821,889 QA pairs
- On average, 5.2 questions per dialog

Popular Datasets

- GuessWhat?!
- VisDial

Crowdsourcing Task

Caption: A sink and toilet in a small room.

You have to ASK questions about the image.

Fellow Turker connected.
Now you can send messages.

1.You:
is this a bathroom ?

1.Fellow Turker:
yes, it's a bathroom

2.You:
what color is the room ?

Message

SEND

(a) What the 'questioner' sees.

Caption: A sink and toilet in a small room.

You have to ANSWER questions about the image.

Fellow Turker connected.
Now you can send messages.

1.Fellow Turker:
is this a bathroom ?

1.You:
yes, it's a bathroom

2.Fellow Turker:
what color is the room ?

2.You:
it looks cream colored

Message

SEND

(b) What the 'answerer' sees.



Caption:
A sink and toilet in a small room.

Q3: can you see anything else ?
A3: there is a shelf with items on it
Q4: is anyone in the room ?
A4: nobody is in the room
Q5: can you see on the outside ?
A5: no, it is only inside
Q6: what color is the sink ?
A6: the sink is white
Q7: is the room clean ?
A7: it is very clean
Q8: is the toilet facing the sink ?
A8: yes the toilet is facing the sink
Q9: can you see a door ?
A9: yes, I can see the door
Q10 what color is the door ?
A10 the door is tan colored

(c) Example dialog from our VisDial dataset.

Workers can end a conversation after 20 messages are exchanged (10 question-answer pairs)

Asking Crowdsourcing Interface

Caption: A man, wearing goggles and a backpack on skis pulls a girl on skis behind him.

You have to ASK Questions about the image.

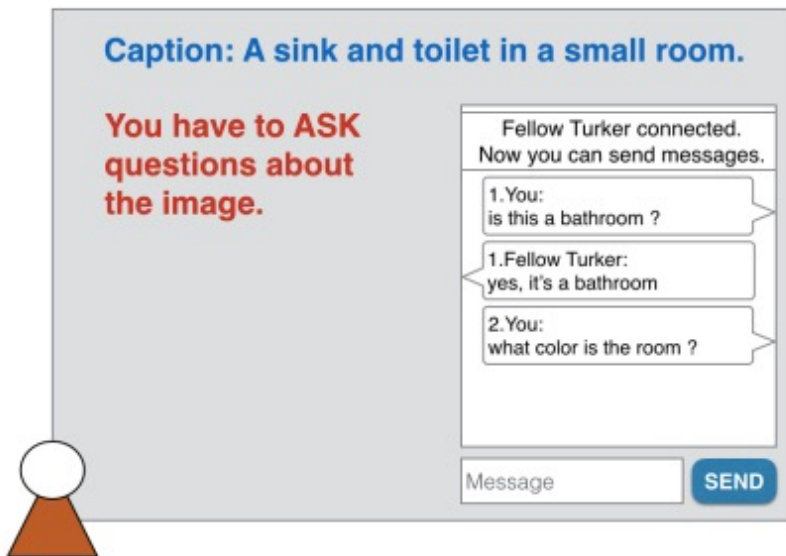
Fellow Turker connected. Now you can send messages

Type Message Here:

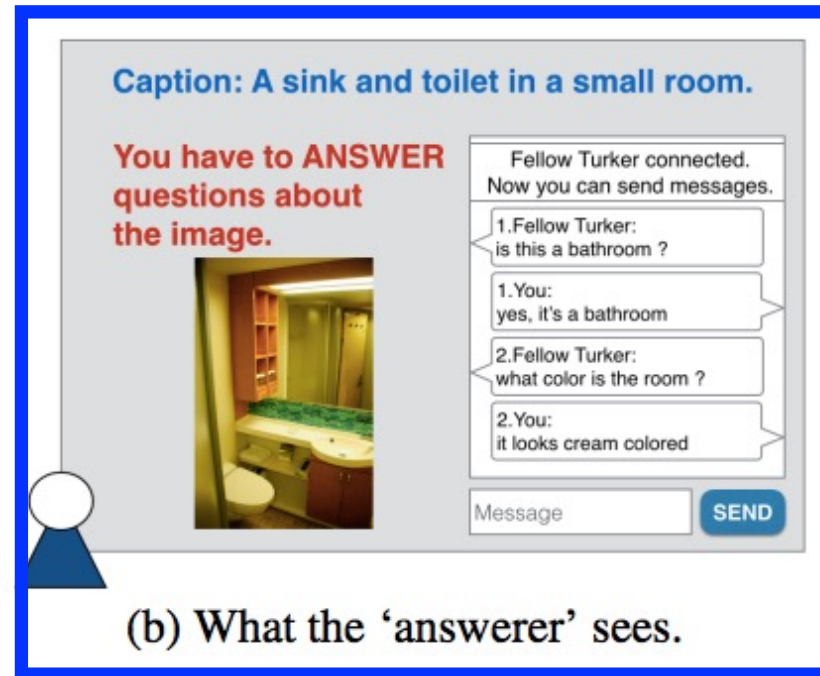
What is the benefit of not showing the image?

- No visual priming; questions help to create a mental model

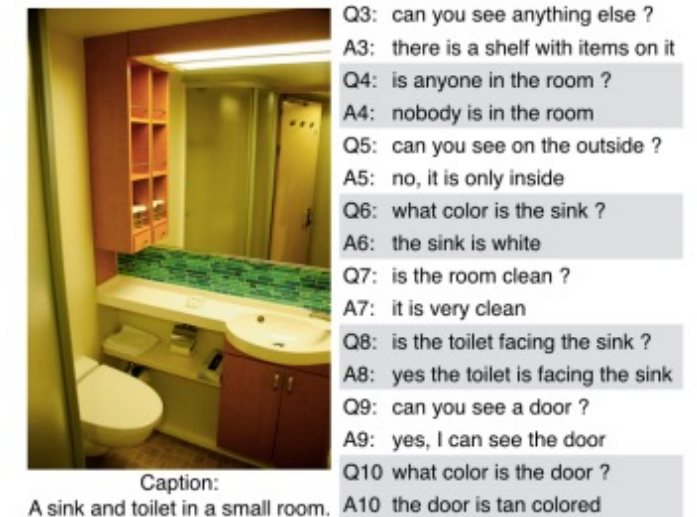
Crowdsourcing Task



(a) What the 'questioner' sees.



(b) What the 'answerer' sees.



(c) Example dialog from our VisDial dataset.

Workers can end a conversation after 20 messages are exchanged (10 question-answer pairs)

Answering Crowdsourcing Interface

Caption: A man, wearing goggles and a backpack on skis pulls a girl on skis behind him.

You have to ANSWER questions about the image.



Fellow Turker connected. Now you can send messages

Type Message Here:

Message

Send

End Conversation And Finish Hit

Crowdsourcing Instructions

Live Question/Answering about an Image.

▼ Instructions

In this task, you will be talking to a fellow Turker. You will either be asking questions or answering questions about an image. You will be given more specific instructions once you are connected to a fellow Turker.

Stay tuned. A message and a beep will notify you when you have been connected with a fellow Turker.

Please keep the following in mind while chatting with your fellow Turker:

- 1 Please directly start the conversation. Do not make small talk.
- 2 Please do not write potentially offensive messages.
- 3 Please do not have conversations about something other than the image. Just either ask questions, or answer questions about an image (depending on your role).
- 4 Please do not use chat/IM language (e.g, "r8" instead of "right"). Please use professional and grammatically correct English.
- 5 **Please have a natural conversation. Unnatural sounding conversation including awkward messages and long silences will be rejected.**
- 6 Please note that you are expected to complete and submit the hit in one go (once you have been connected with a partner). You cannot resume hits.
- 7 **If you see someone who isn't performing HITs as per instructions or is idle for long, do let us know. We'll make sure we keep a close watch on their work and reject it if they have a track record of not doing HITs properly or wasting too much time. Make sure you include a snippet of the conversation and your role (questioner or answerer) in your message to us, so we can look up who the other worker was.**
- 8 **Do not wait for your partner to disconnect to be able to type in responses quickly, or your work will be rejected.**

Please complete one hit before proceeding to the other. Please don't open multiple tabs, you cannot chat with yourself.

VisDial Statistics

- ~140,000 images (from COCO)
- ~2.4M QA pairs with 10 QA pairs per image

Popular Datasets

- GuessWhat?!
- VisDial

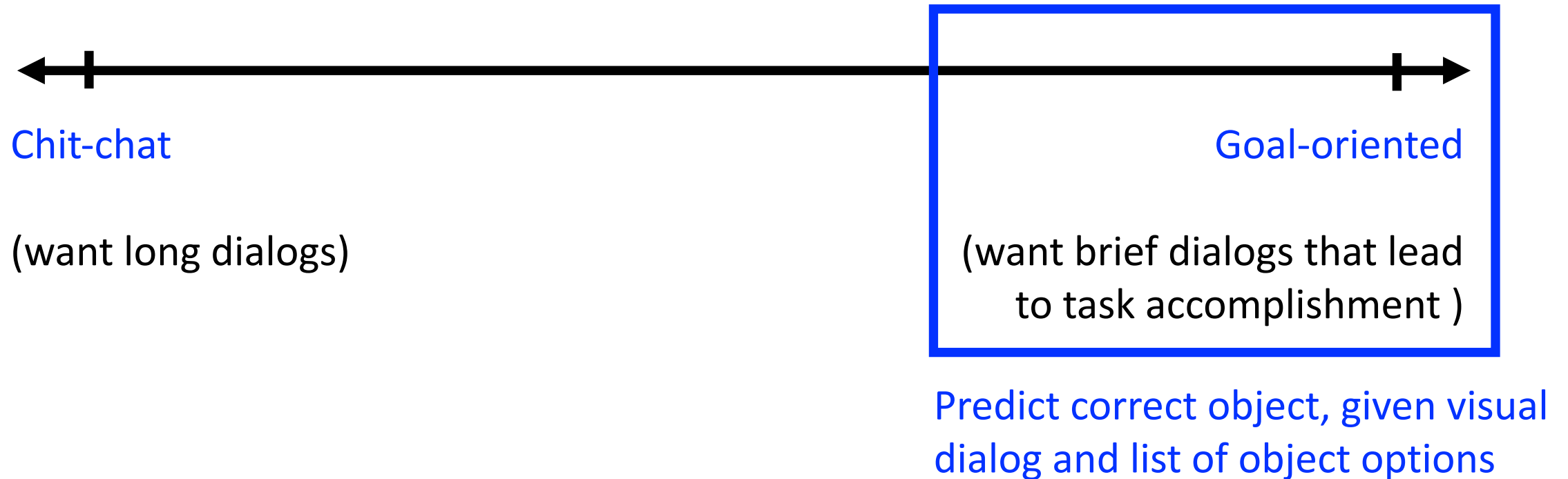
Popular Datasets: VisDial and GuessWhat?!

- What are biases of these datasets, and what might be the impact of such biases on models trained and evaluated on these datasets?

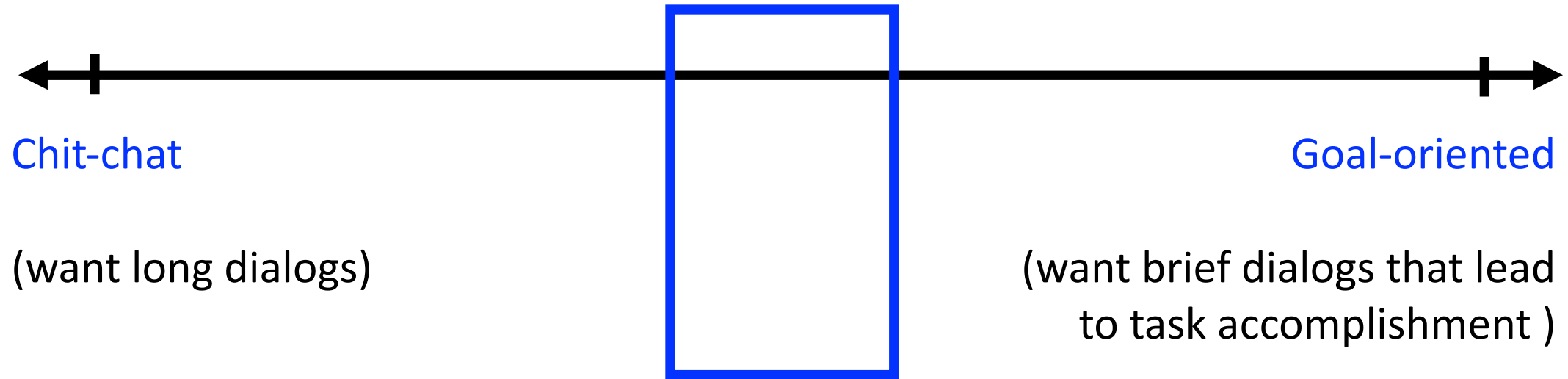
Today's Topics

- Visual dialog applications
- Visual dialog dataset
- **Visual dialog evaluation**
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial

GuessWhat?! Evaluation Metric



VisDial Evaluation Metric



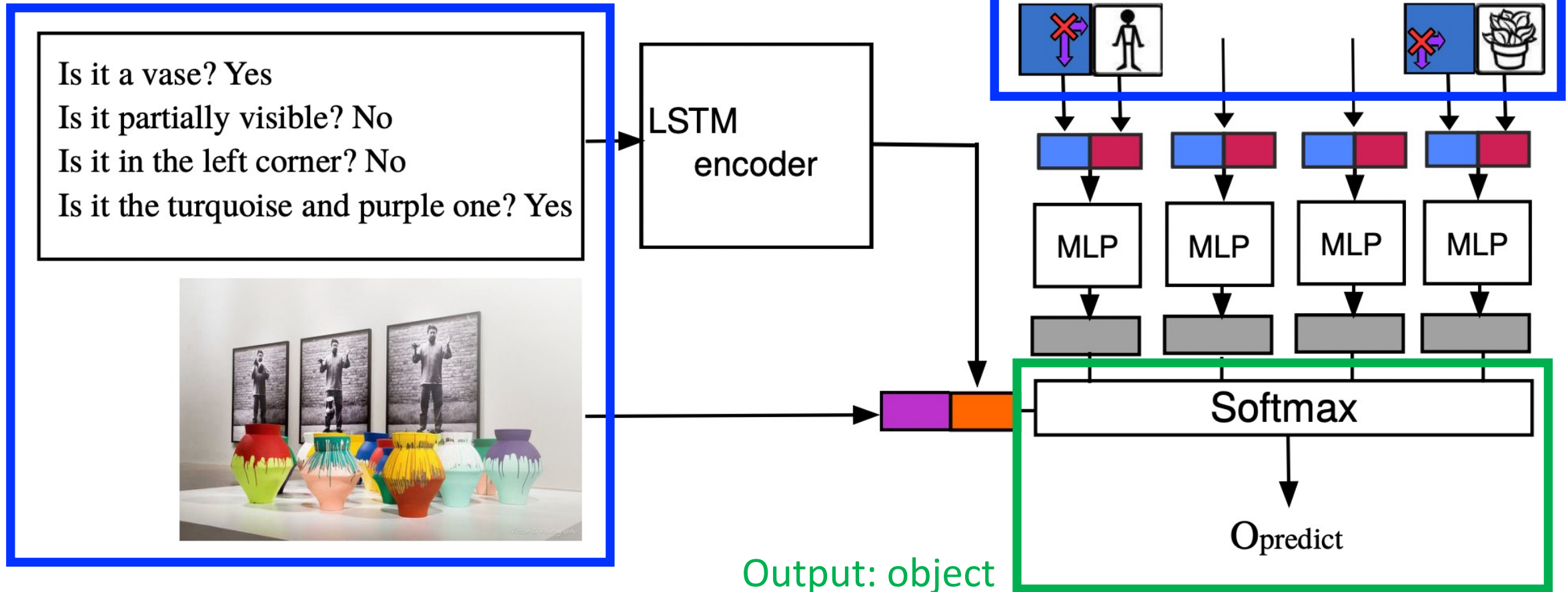
Evaluate for each new QA pair the predicted ranking of answers using retrieval metrics (e.g., recall@k); 100 candidate answers are provided with the visual dialog

Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- **Mainstream 2017 challenges: baseline approaches**
- LTMI: Transformer approach
- Latex tutorial

GuessWhat?! Baseline Model

Input: dialog history, image, and...



GuessWhat?! Baseline Model

Question representation: hidden state after feeding all images and questions as a sequence

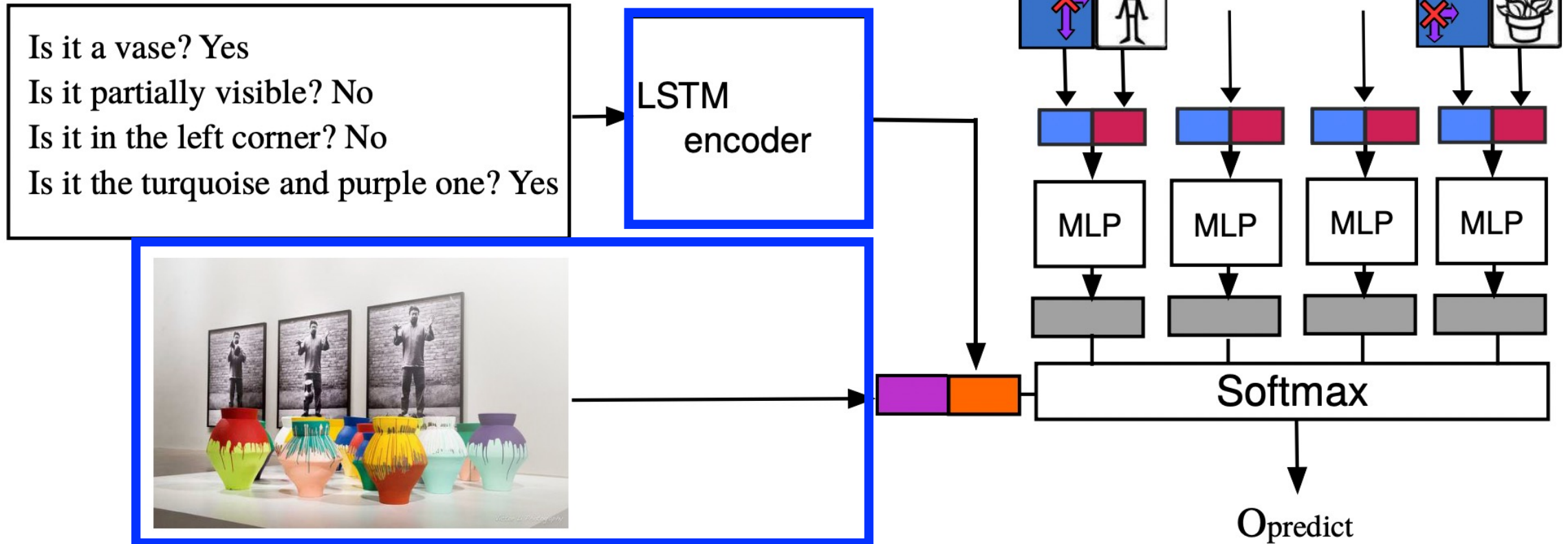
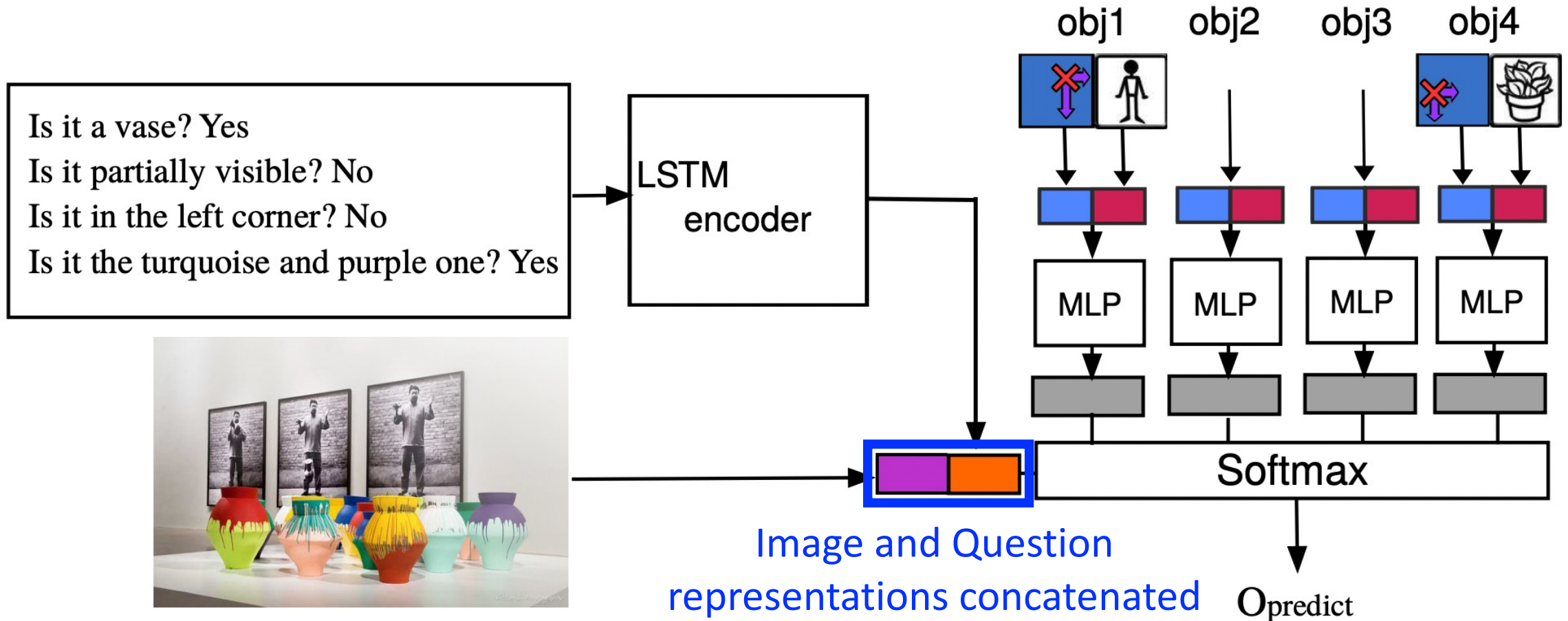


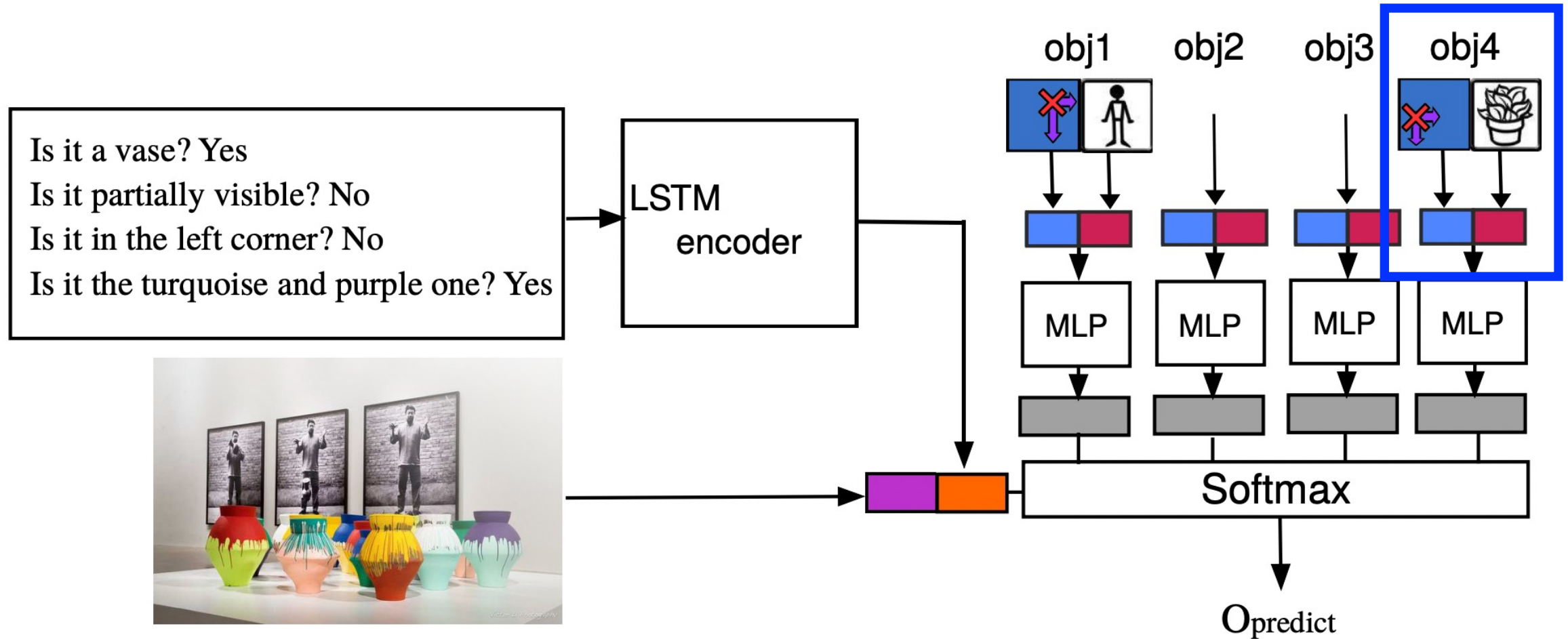
Image representation: FC8 features from VGG16

GuessWhat?! Baseline Model

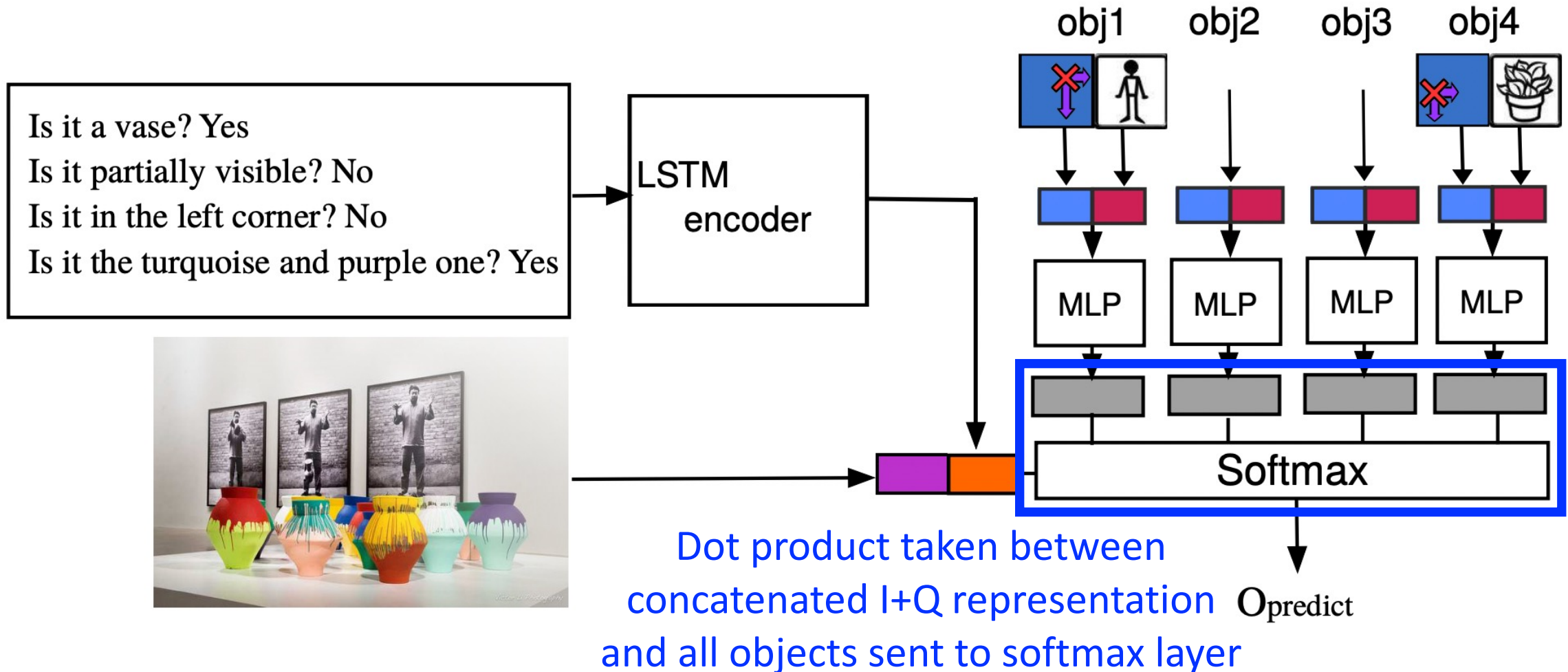


GuessWhat?! Baseline Model

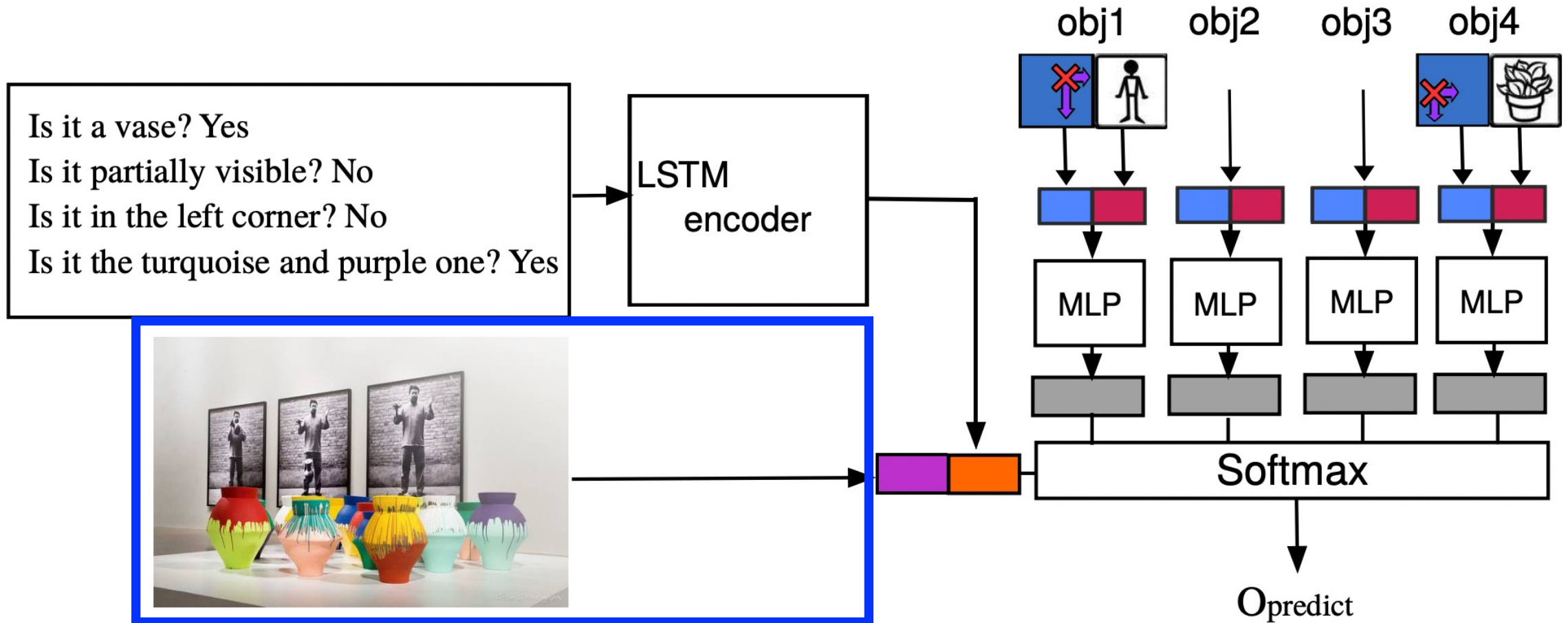
Object representation:
category and spatial features



GuessWhat?! Baseline Model

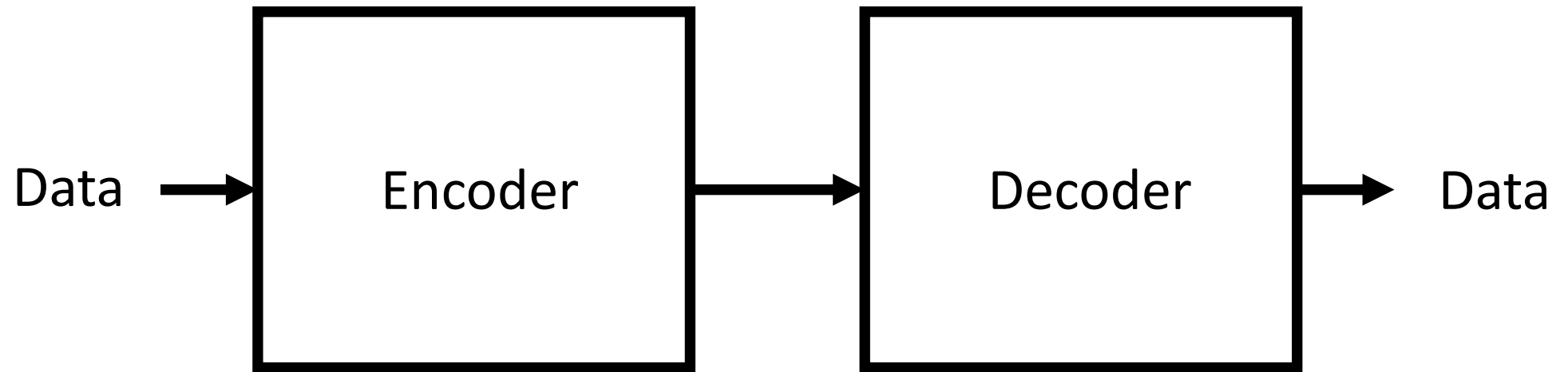


GuessWhat?! Experimental Results

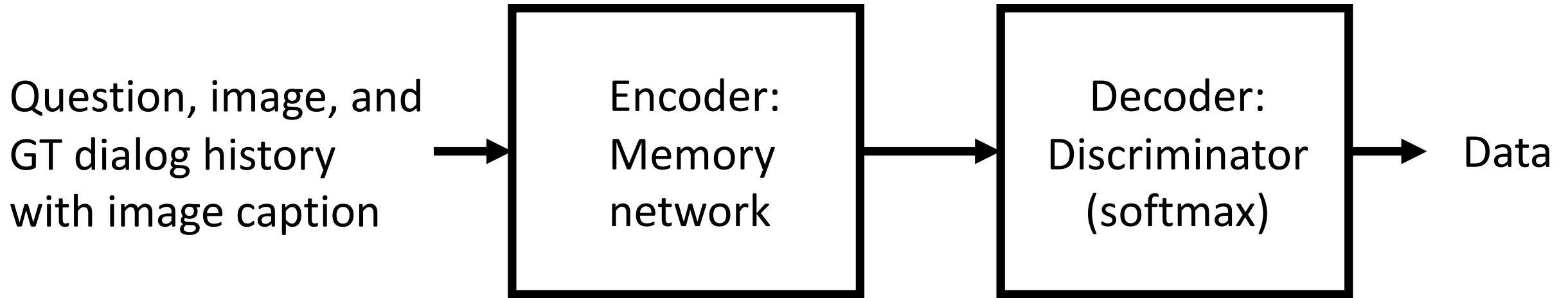


Ablation study that excluded image features revealed they are not helpful for prediction

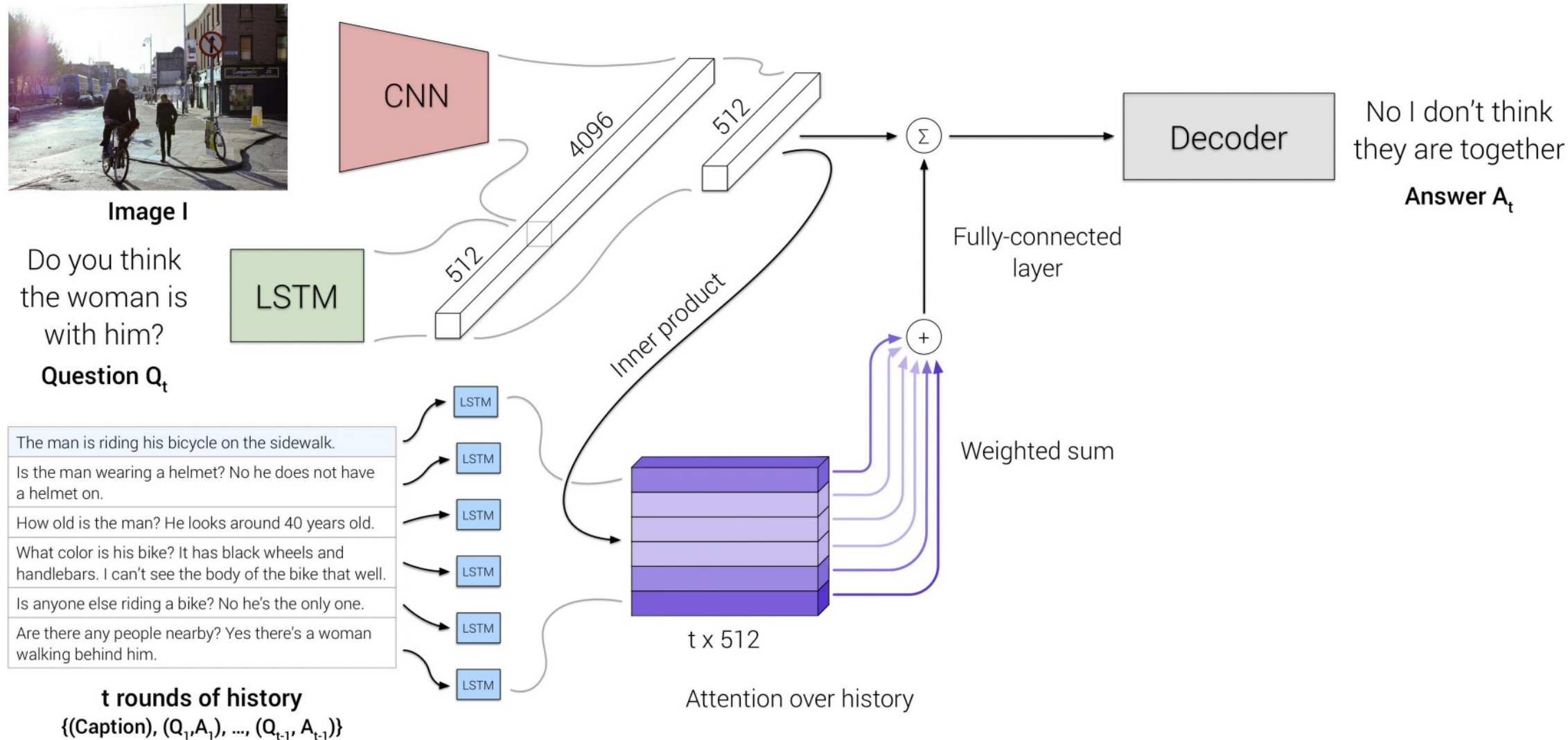
VisDial Baseline Model



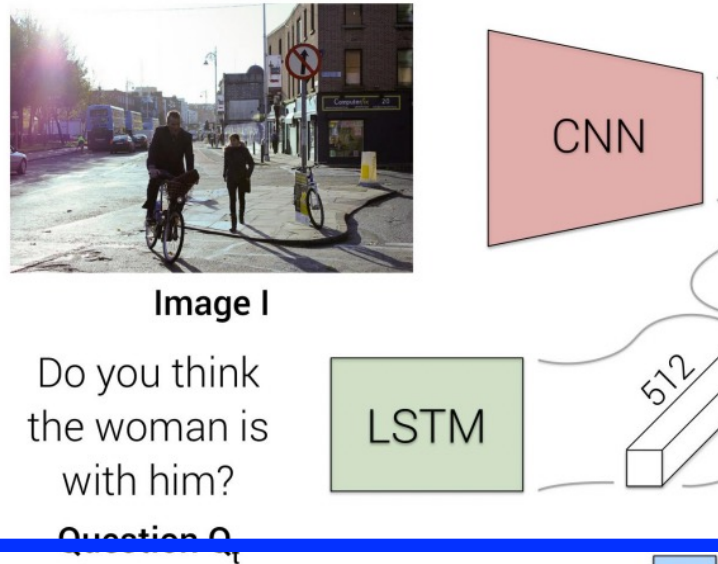
VisDial Baseline Model



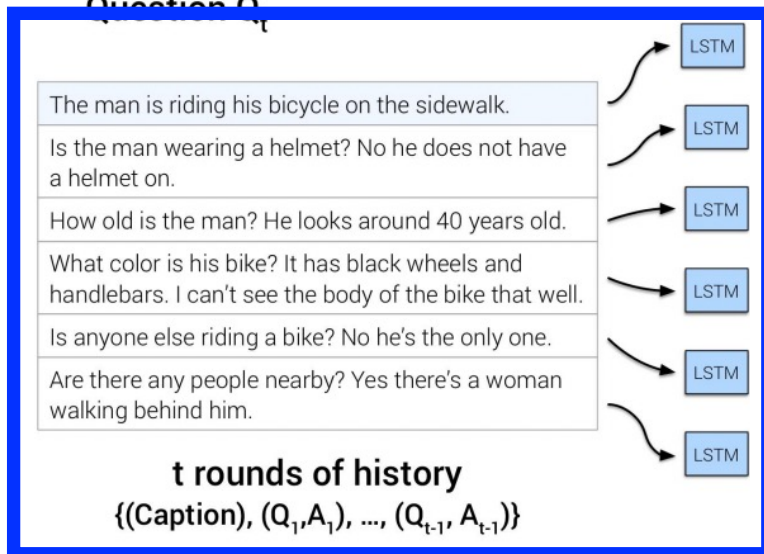
VisDial Baseline Model



VisDial Baseline Model



- How many rounds of history are observed in this example as context for predicting the answer to the next question?



Dialog history: each caption and QA pair encoded by the same LSTM into hidden representations

VisDial Baseline Model

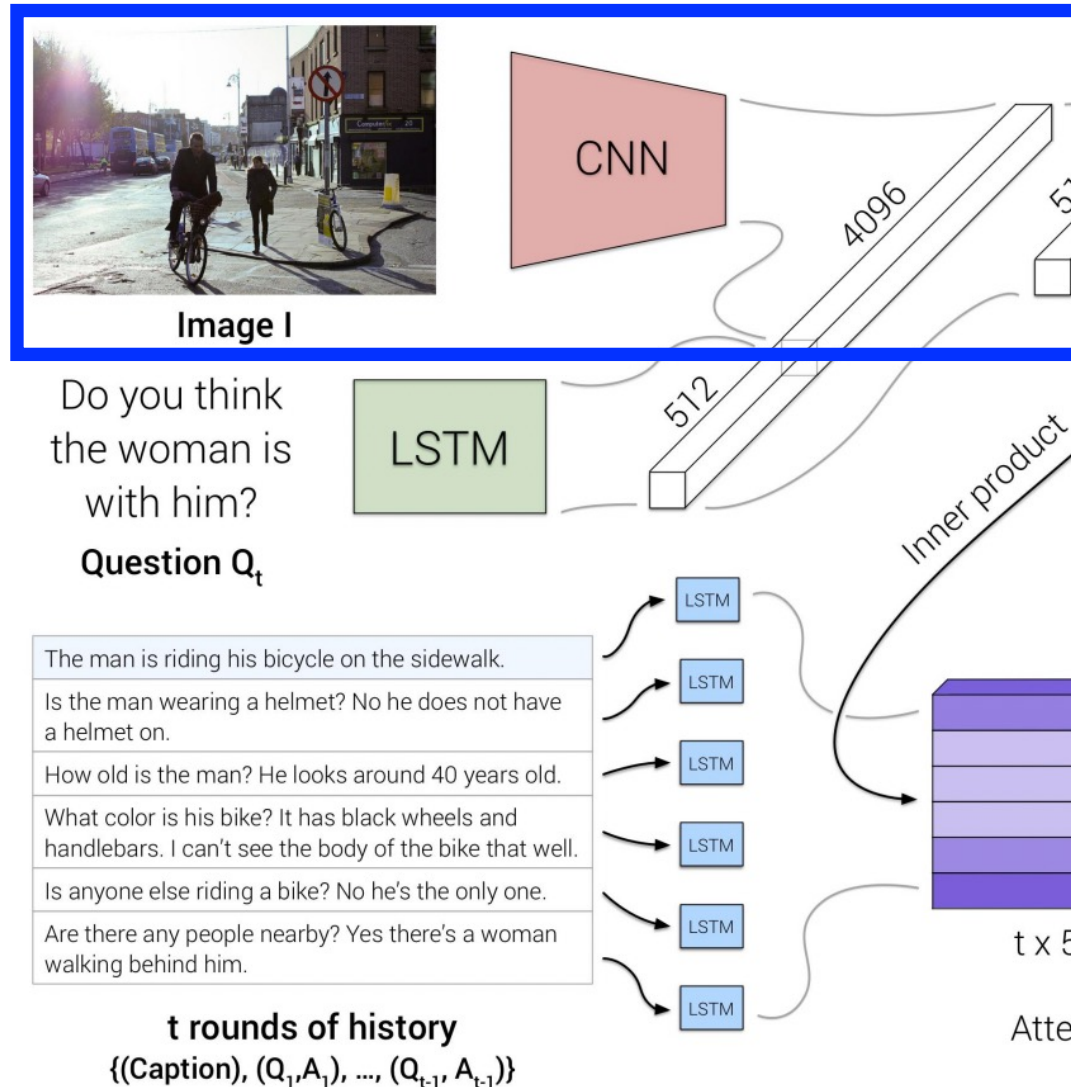
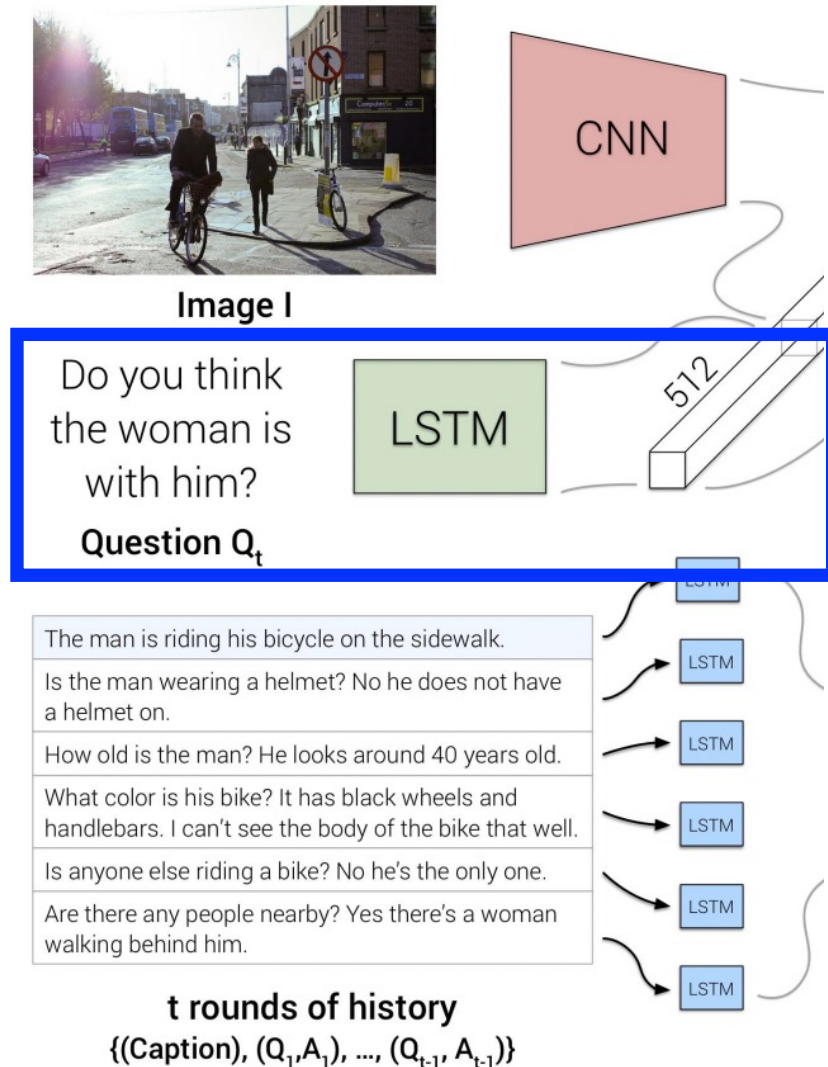


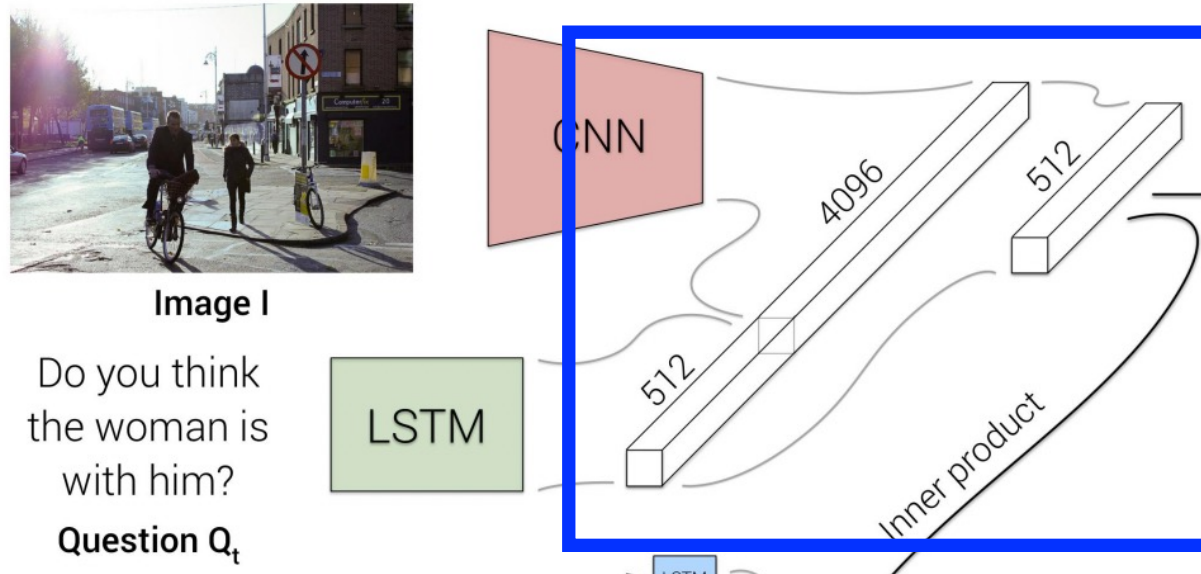
Image representation: features from second last layer of VGG-16

VisDial Baseline Model



Question representation: last hidden representation from input question

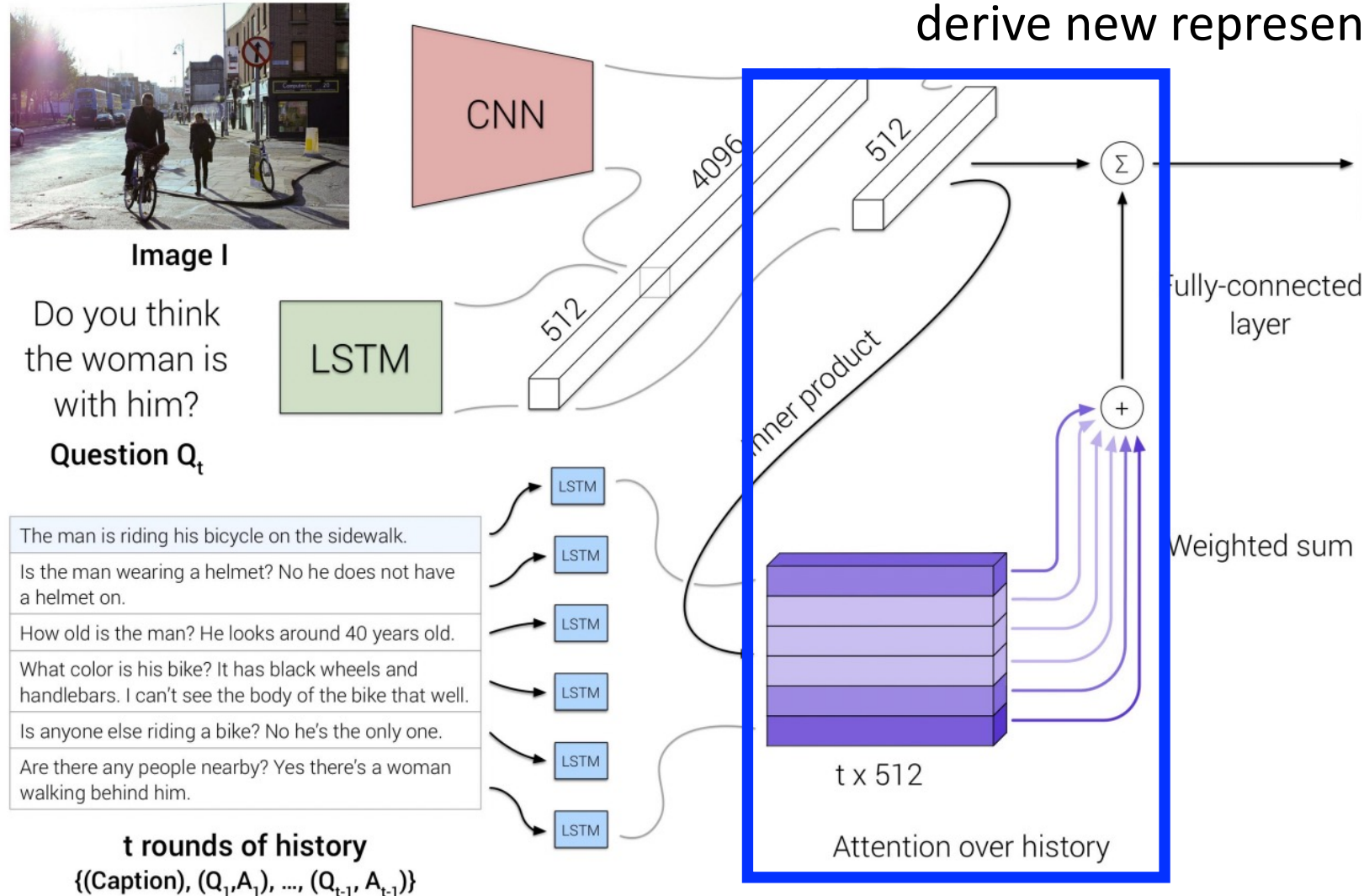
VisDial Baseline Model



Cross-modality representation:
concatenated features followed
by fully-connected layer with
tanh activation function

VisDial Baseline Model

New encoder representation: attention weight of each “fact” (i.e., caption, QA pair) for the multimodal query used to derive new representation



VisDial Baseline Model



Image I

Do you think
the woman is
with him?

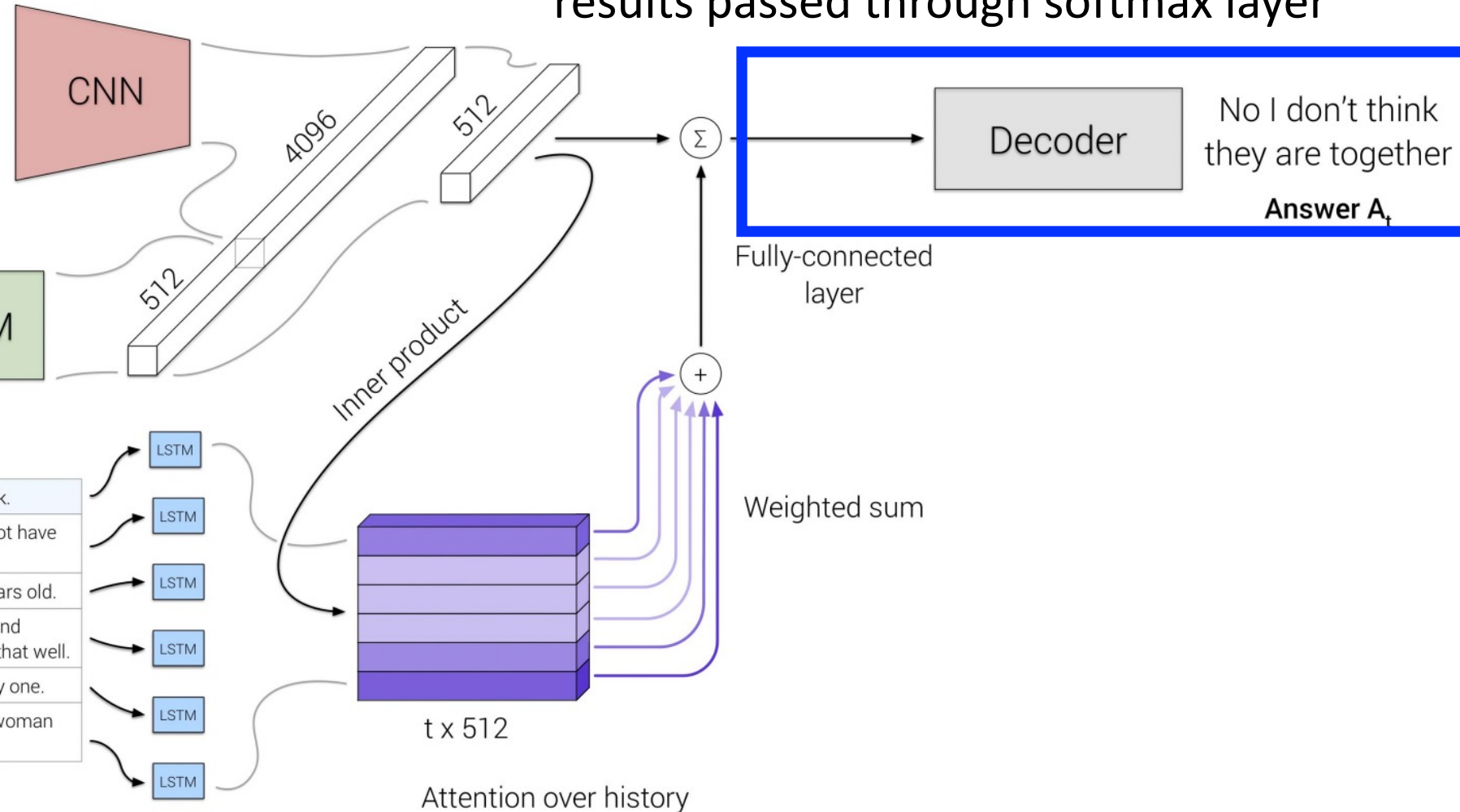
Question Q_t



The man is riding his bicycle on the sidewalk.
Is the man wearing a helmet? No he does not have a helmet on.
How old is the man? He looks around 40 years old.
What color is his bike? It has black wheels and handlebars. I can't see the body of the bike that well.
Is anyone else riding a bike? No he's the only one.
Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history

$\{(Caption), (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$



Prediction: similarity between input encoding and each of 100 given candidate answers measured using dot product; all results passed through softmax layer

Qualitative Results

(color intensity indicates attention to the fact)



Is anyone on bus?

A large yellow bus parked in some grass.

Are there any black stripes? Yes 3 black stripes

Is there any writing? Yes it says "moon farm day camp"

Is grass well-maintained? No it's all weeds



What color is his board?

A surfer wiping out on an ocean wave.

Is it man or woman? Man

Are they wearing wetsuit? No



Is it fairly close up shot?

A nice bird standing on a bench.

Gazing at? Camera I think

Can you tell what kind of bird it is?
No it's bright red bird with black face and red beak

Is it tiny bird? Yes

What sort of area is this in? Looks like it could be back deck

Qualitative Results



Are there people on carriage?

A street scene with a horse and carriage.

Is it real? Yes

What color is horse? Dark brown

What color is carriage? Red



What color are kites?

A lot of people stand around flying kites in a park.

Are these people children? It looks like a mixture of families

Is this field trip you think? Just family outing

Is there lot of grass? Yes

Are there lot of trees? No

Any vehicles around? No



Can you see street signs?

The computer on the desk shows an image of a car.

What color is car? White

Do you know make? Volkswagen

Are there people? Probably driving car

Is it in office? It's close up of desk so can't tell

Do you see desk? Yes

Is it laptop? No, desktop

What color is computer? You can't see actual computer just screen and keyboard

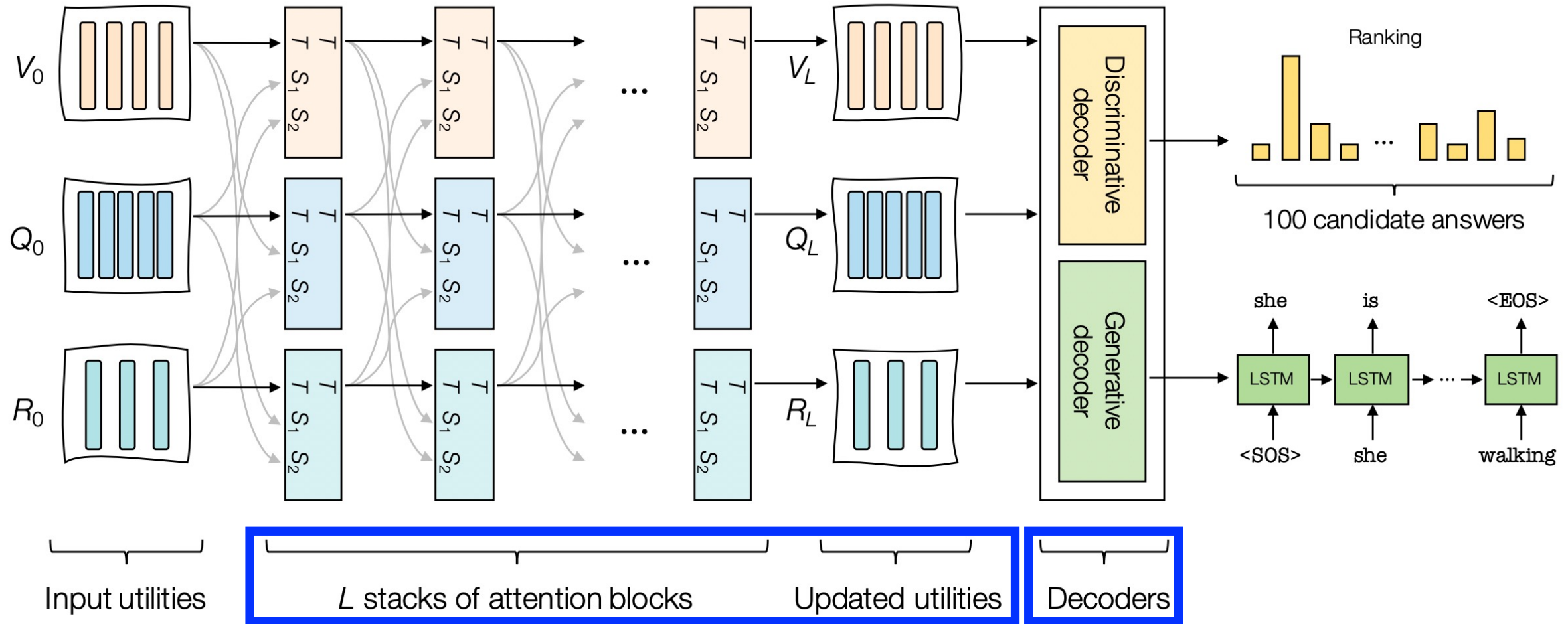
Can you see brand? It's Mac

Is picture of car taken outside? Yes

Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial

Architecture: Encoder-Decoder

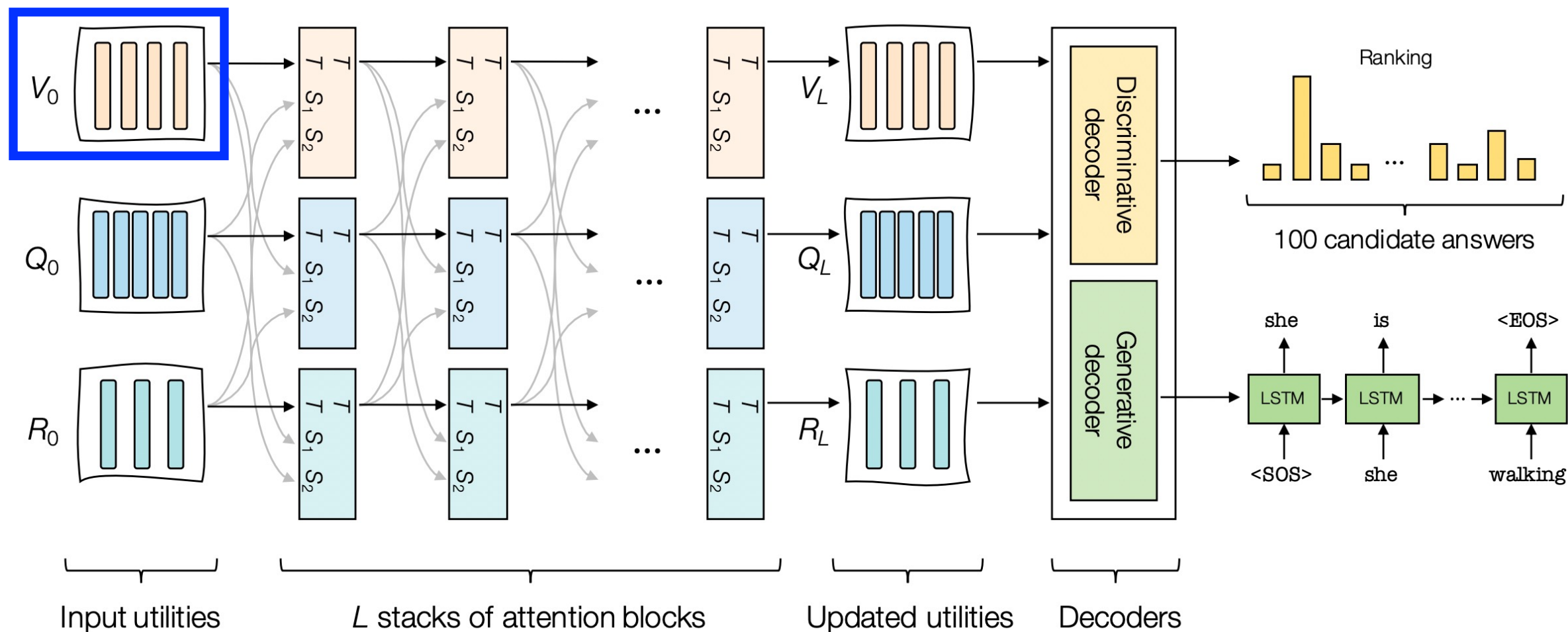


Encoder

Nguyen et al. Efficient Attention Mechanism for Visual Dialog that can Handle All the Interactions between Multiple Inputs. ECCV 2020.

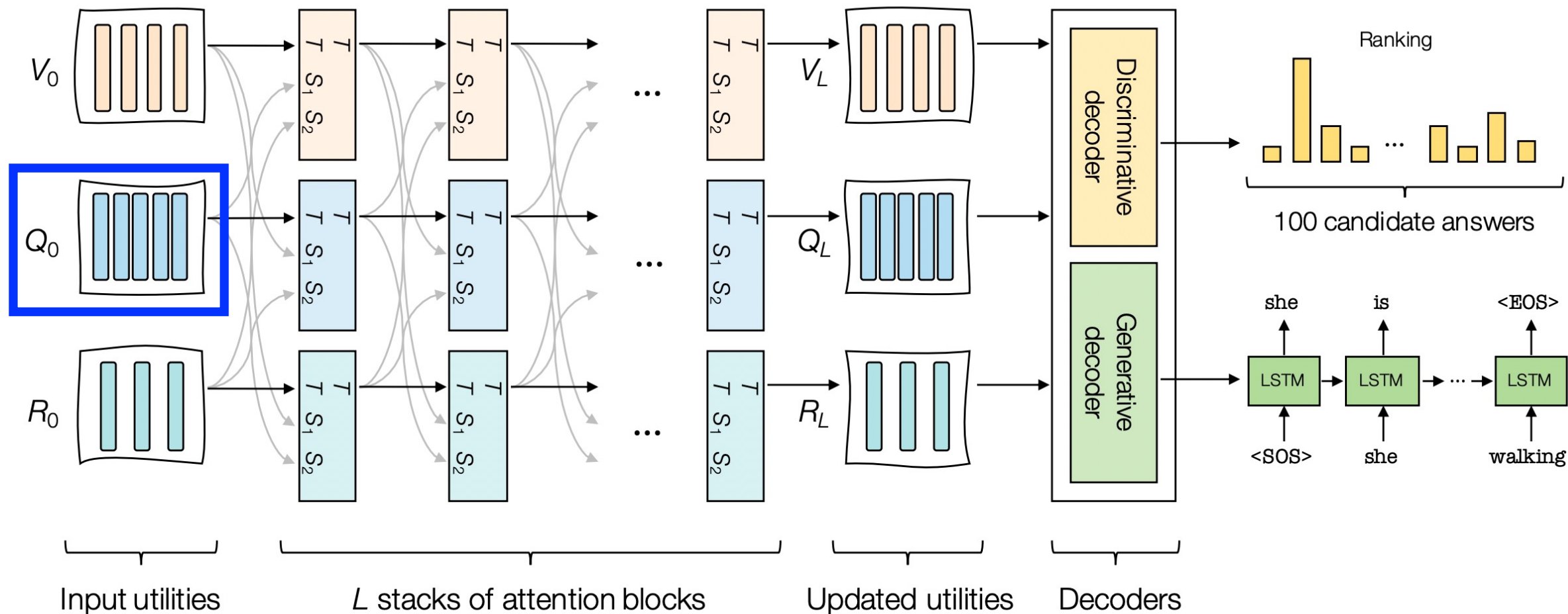
Architecture: Input

Image: 100 object features from Faster R-CNN that fuse image region features and BB positions



Architecture: Input

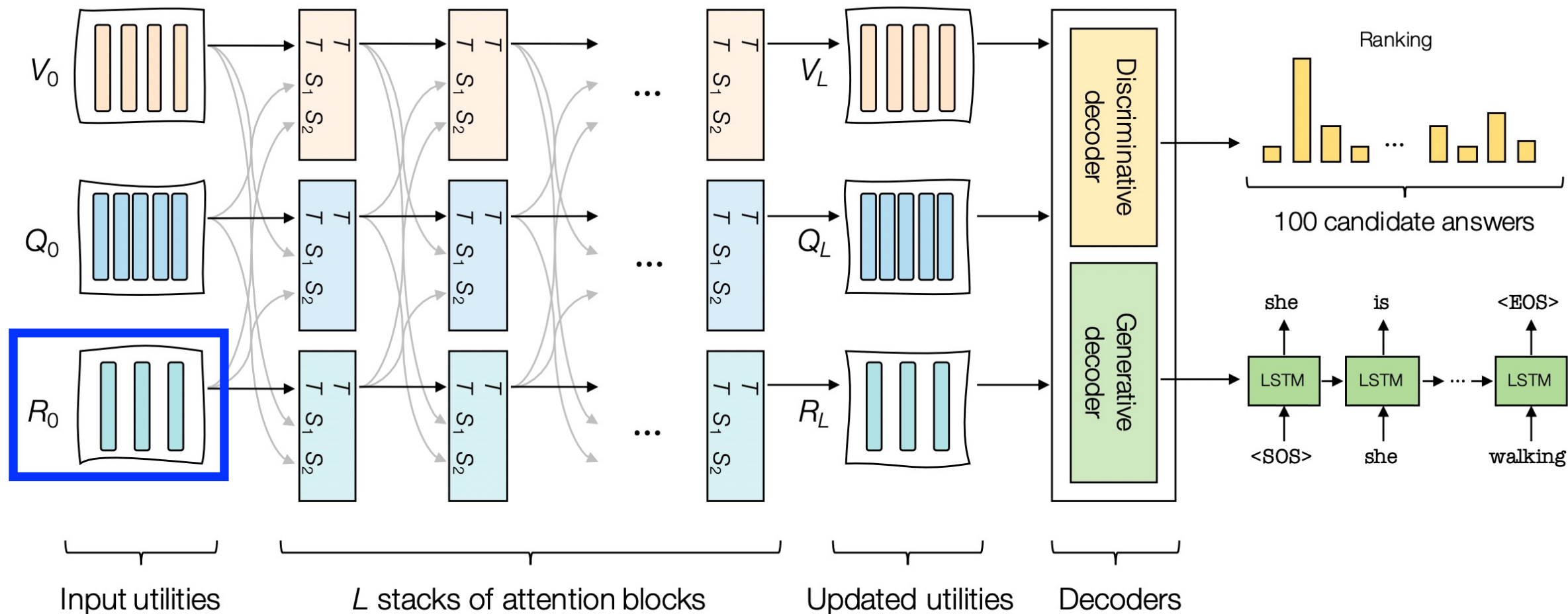
Question: word features based on GloVe embeddings per word



Nguyen et al. Efficient Attention Mechanism for Visual Dialog that can Handle All the Interactions between Multiple Inputs. ECCV 2020.

Architecture: Input

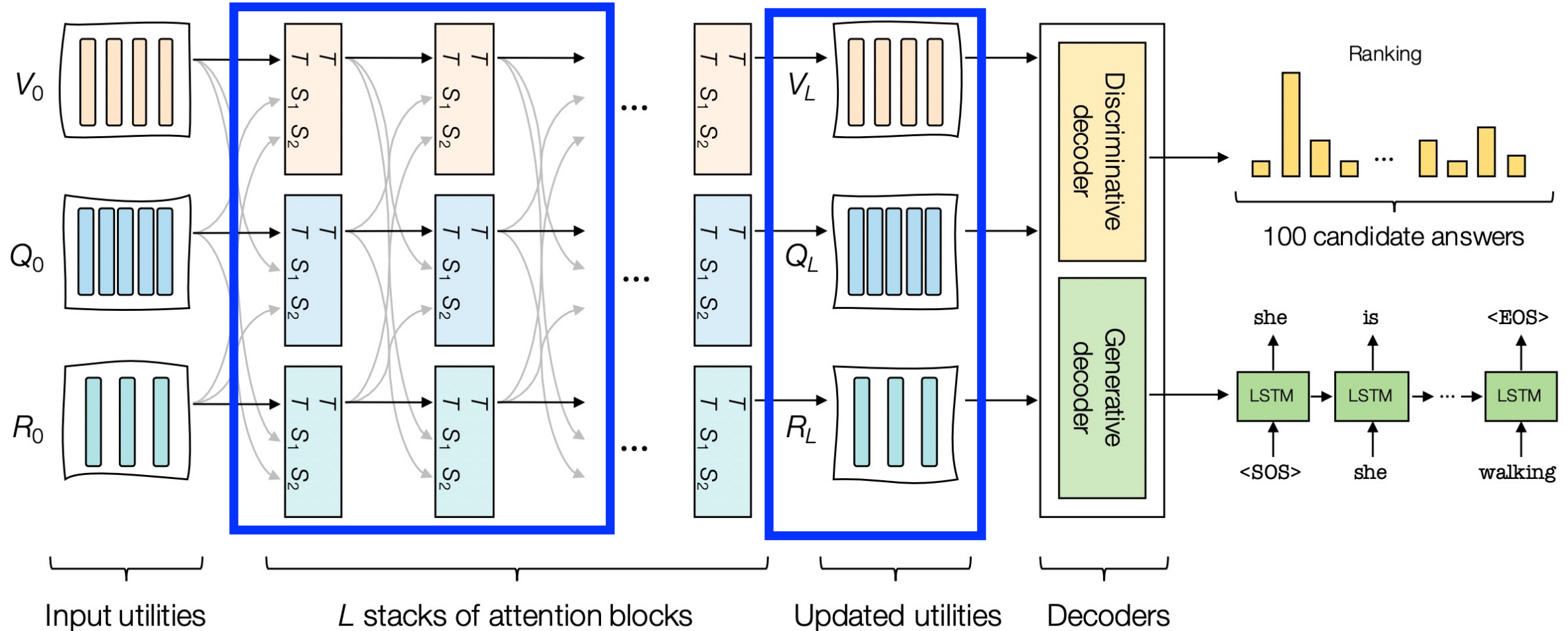
Dialog history: each round (caption, QA pairs) embedded to same length with truncation/padding



Nguyen et al. Efficient Attention Mechanism for Visual Dialog that can Handle All the Interactions between Multiple Inputs. ECCV 2020.

Architecture: Encoder-Decoder

New representations of inputs learned by transformer

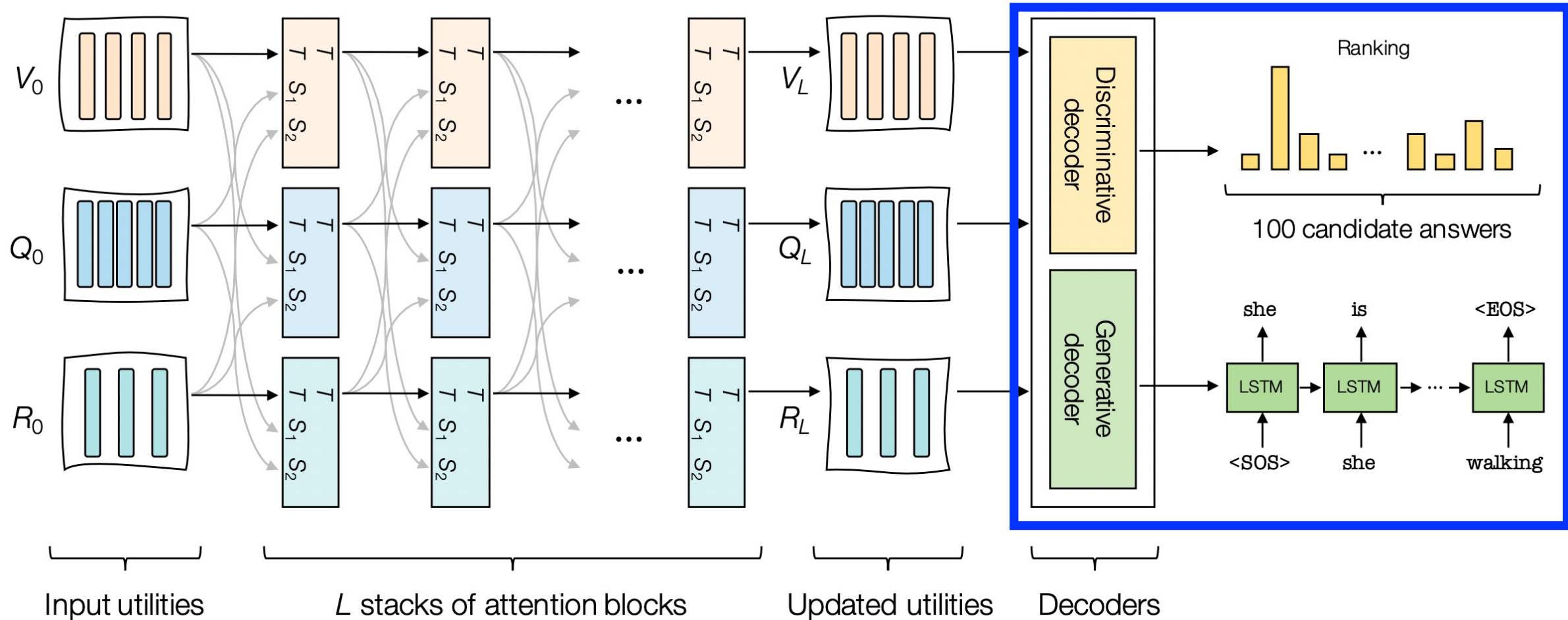


Light-weight Transformer for Many Inputs

Nguyen et al. Efficient Attention Mechanism for Visual Dialog that can Handle All the Interactions between Multiple Inputs. ECCV 2020.

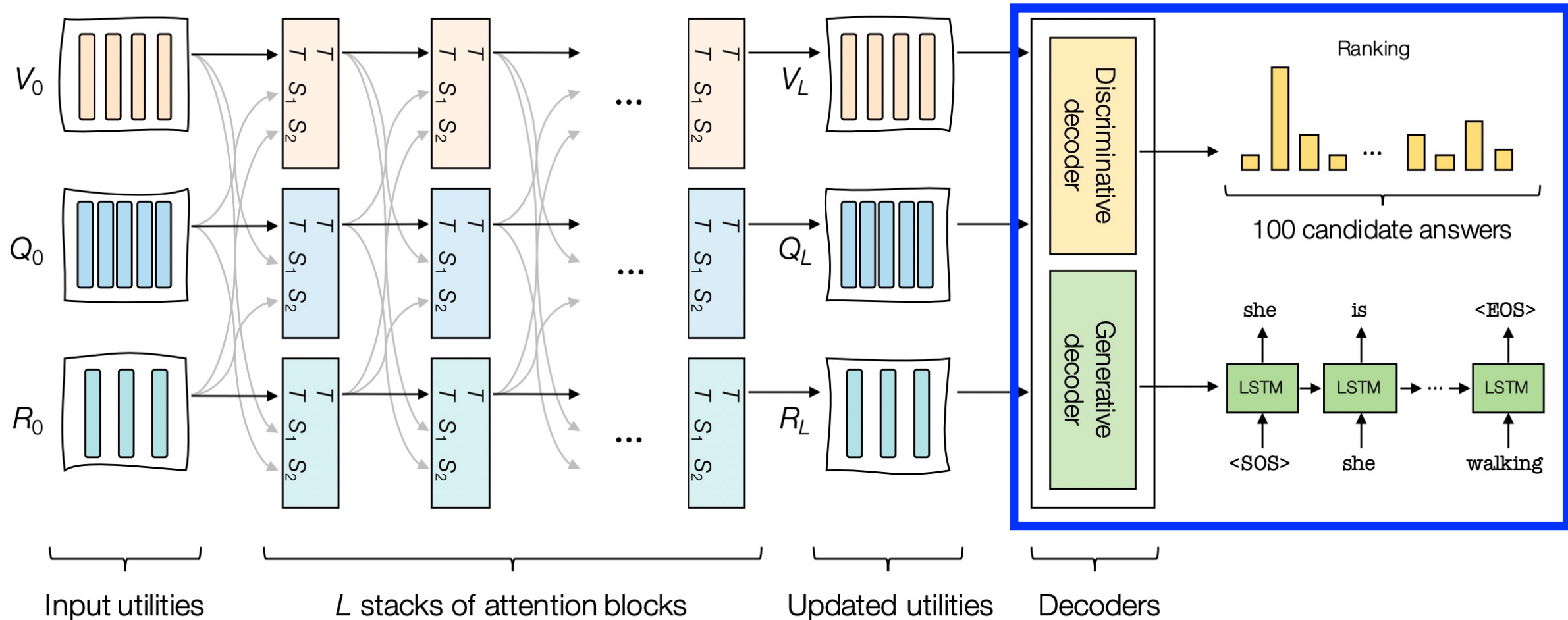
Architecture: Multi-Task Learning

Objective: minimize sum of losses from both the discriminator and generator



Architecture: Inference

Either decoder can be used to generate final prediction (latter slightly better on VisDial)



Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial

Today's Topics

- Visual dialog applications
- Visual dialog dataset
- Visual dialog evaluation
- Mainstream 2017 challenges: baseline approaches
- LTMI: Transformer approach
- Latex tutorial



The End