ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

AAAI2020: Visual Dialogue State Tracking for Question Generation

ECCV2020: Guessing State Tracking for Visual Dialogue

Wei Pang, Xiaojie Wang

2019.03～2019.11

School of Artificial Intelligence

Beijing University of Posts and Telecommunications, China

Outline

- Introduction
- GuessWhat?!
- Question Generation & Guesser(2019.03-11)
- Experiments & Conclusions
- GuessWhich & Visual Dialog(VisDial)(2020～)

Outline

- Introduction
- GuessWhat?!
- Question Generation & Guesser(2019.03-11)
- Experiments & Conclusions
- GuessWhich & Visual Dialog(VisDial)(2020～)

Introduction

- Dialogue is a major way of human communication
  - Daily chatting, Customer service, Negotiations …
- Dialogues always occur in some contexts
- Visual scenes are one major type of context: Visual Dialogue

Boy: Hello!
Girl: Hi
Boy: What a sunny day!
Girl: um, it's a good day, but a little hot.
Boy: You look beautiful with the red skirt!
Girl: thank you…

Introduction

- Build Visual Dialogue agents is an interesting AI topic
- Some Visual Dialogue Tasks
  - GuessWhat?! (de Vries, et al. CVPR2017)
  - Visual Dialog(VisDial) (A. Das, et al. CVPR2017)
  - GuessWhich?! (P. Chattopadhyay, et al. AAAI2017)
  - Image-Grounded Conversation (Mostafazadeh, et al. IJCNLP2017)
  - Multimodal Dialogue (MMD) (A. Saha, et al. AAAI2018)
  - MNIST Counting Dialog (S. Lee, et al. NeurIPS2018)
  - CoDraw (Jin-Hwa Kim, et al., ACL2019)
  - VGDS (Le, et al., ACL 2019)

GuessWhich AAAI2017

- GuessWhich is a cooperative image-guessing game between two agents: Q-BOT and A-BOT
  - The goal of GuessWhich is to figure out a correct answer out of 9,628 test images by asking a sequence of questions.
  - Abot see the target image, which is unknown to Qbot
  - Both Qbot and Abot see a brief description of the target image.

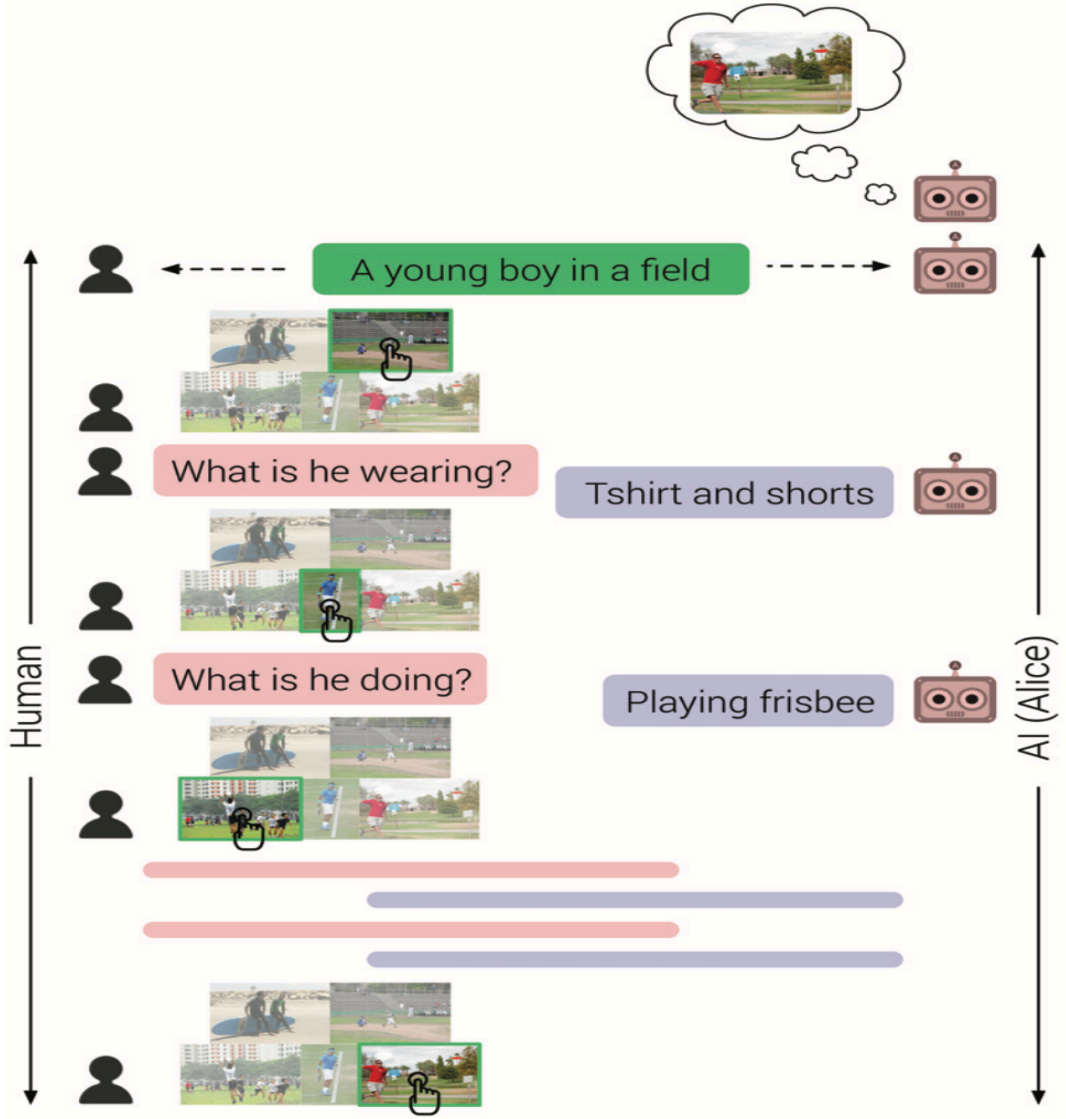


Image-Grounded Conversation CVPR2017

- A naturally-occurring Image-Grounded Conversation.
  - ☐ Event-centric conversations on image
  - ☐ event-driven utterance

User1: My son is ahead and surprised!
User2: Did he end up winning the race?
User1: Yes he won, he can't believe it!

Visual Context			
Textual Context	This wasn't the way I imagined my day starting.	I checked out the protest yesterday.	A terrible storm destroyed my house!
Question	do you think this happened on the highway?	Do you think America can ever overcome its racial divide?	OH NO, what are you going to do?
Response	Probably not, because I haven't driven anywhere except around town recently.	I can only hope so.	I will go live with my Dad until the insurance company sorts it out.

MMD AAAI2018

Shopping Assistant ChatLog

SHOPPER: show me some espadrilles in beige/blue within $200.

AGENT: Sure. let me just quickly browse through my catalogue

AGENT:

AGENT: What you can see in the fifth item, michael michael kors brings you a weekend look with the wrapped espadrille heel and crisscross styling of these dana flats.

SHOPPER: List more in the upper material as in the 5th image and style as the 3rd and 5th

AGENT:

SHOPPER: Can you list images of the 1st result from all orientations?

AGENT: Image from the front,right and back orientations respectively

SHOPPER: And the 3rd one?

AGENT: Image from the front,right,back and left orientations respectively

SHOPPER: I like the 4th image. Show me something like it but in style as in this image, within my budget.

AGENT:

SHOPPER: Will these espadrilles suit office style?

AGENT: Yes

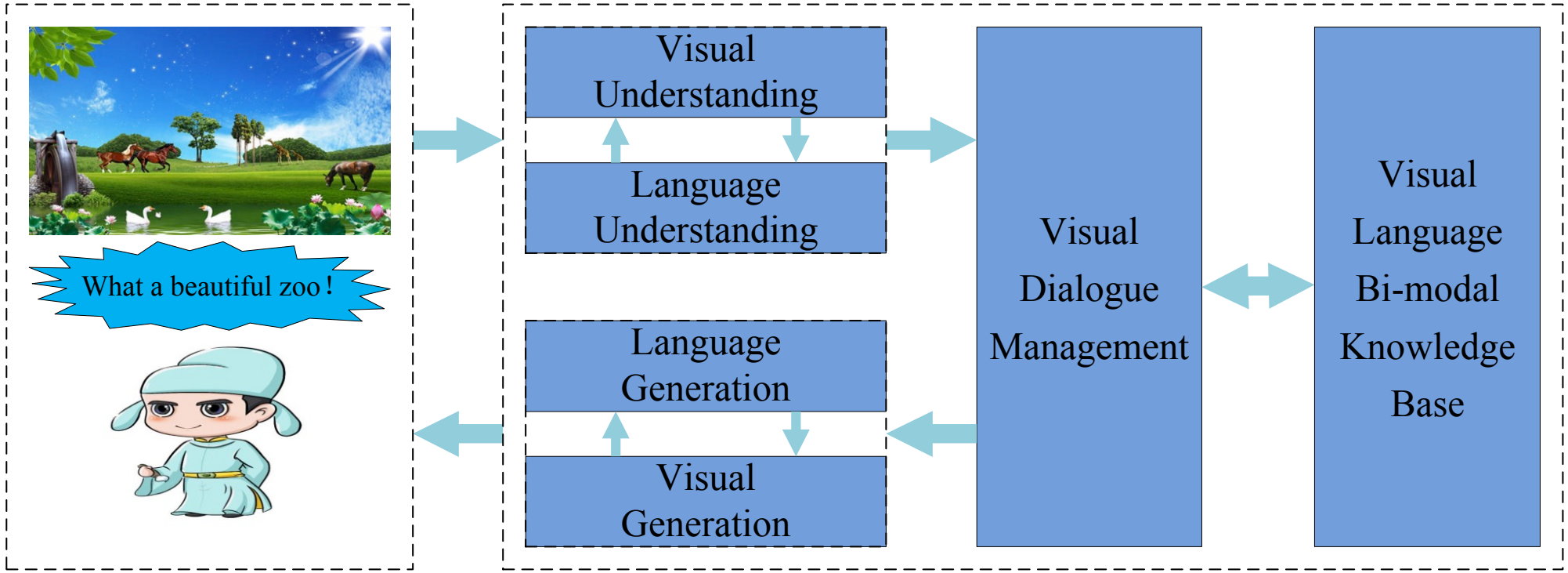
Video-Grounded Dialogue System ACL2019

- Dialog is based on a given video
- Video information
  - a sequence visual images
  - Audio
- Text caption
- 10 rounds dialog between two agents

C: a man is standing in a kitchen putting groceries away. He closes the cabinet when finished, walks over to a table and pulls out a chair and sits down.
S: a man puts away his groceries and then sits at a kitchen table and stares out the window.
Q1: how many people are in the video?
A1: there is just one person
Q2: is there sound to the video?
A2: yes there is audio but no one is talking
...
Q10: is he happy or sad?
A10: he appears to be neutral in expression

Introduction

- Visual Dialogue Agents in general
  - Visual information processing\Language information processing
  - Dialogue policy management\knowledge management
  - Fusion of above



Outline

- Introduction
- GuessWhat?!
- Question Generation & Guesser(2019.03-11)
- Experiments & Conclusions
- GuessWhich & Visual Dialog(VisDial)(2020～)

GuessWhat?! Game

- An object-guessing game between two players around an image
  - An image with several candidate objects in it
  - A target is randomly assigned by Oracle
  - Questioner:
    - Question Generator(QGen):

      asks a series of questions to Oracle
    - Guesser:

      selects an object as target from candidates
  - Oracle: answer the questions with Yes/No

QGen	Oracle
Is it a cow?	Yes
Is it the big cow in the middle?	No
Is the cow on the left	No
On the left?	Yes
First cow near us?	Yes

GuessWhat?! Game

- GuessWhat?! dataset
  - 66,537 images
  - 134,073 objects
  - 160,745 dialogues
  - 821,889 Question Answer pairs

Outline

- Introduction
- GuessWhat?!
- Question Generation & Guesser(2019.03-11)
- Experiments & Conclusions
- GuessWhich & Visual Dialog(VisDial)(2020)

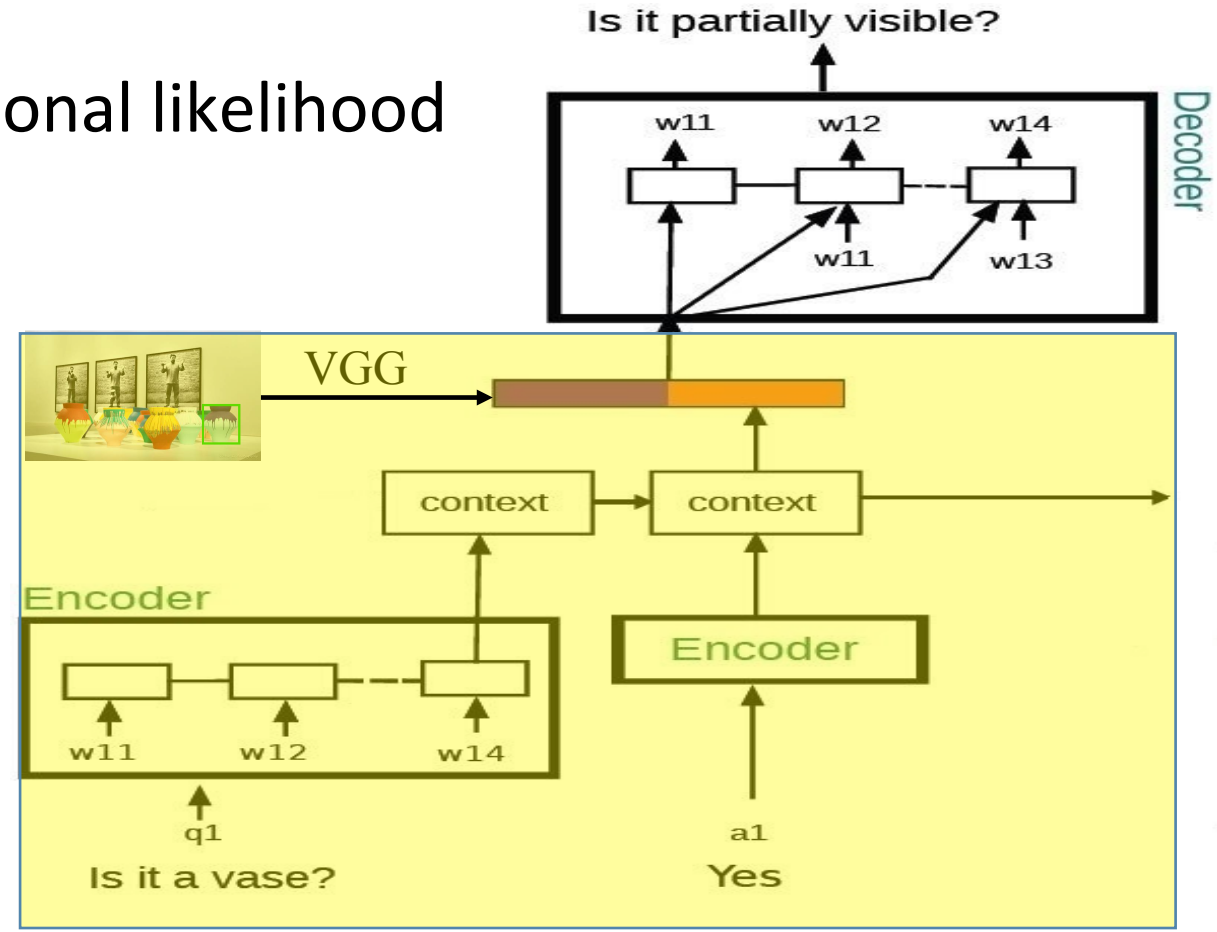
Question Generation(QGen)

- Goal of QGen
  - Given: an image includes several objects
  - Asks a sequence of questions to the Oracle

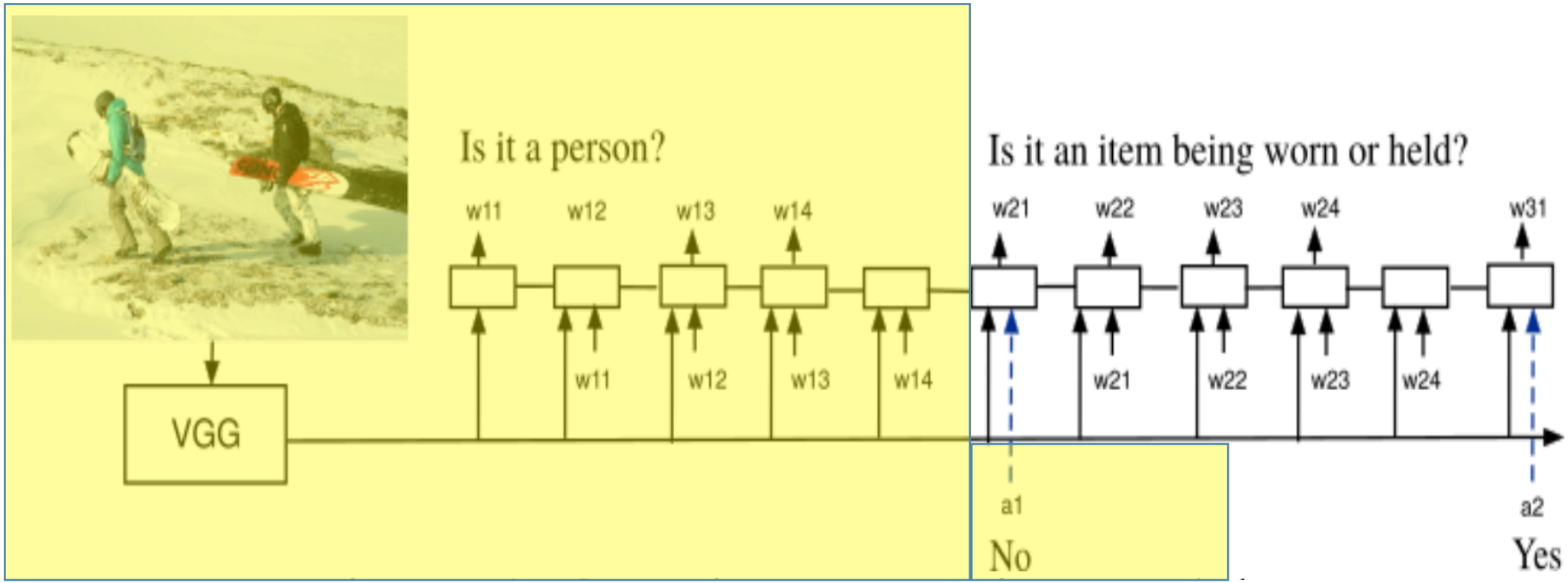
Question Generation(QGen)

- First QGen model: Encoder-Decoder with Supervised Learning (de Vries, CVPR2017)
  - Encoder: Two-layer RNN for Dialogue, VGG for Image, concatenation
  - Decoder: RNN
  - Supervised learning： conditional likelihood

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

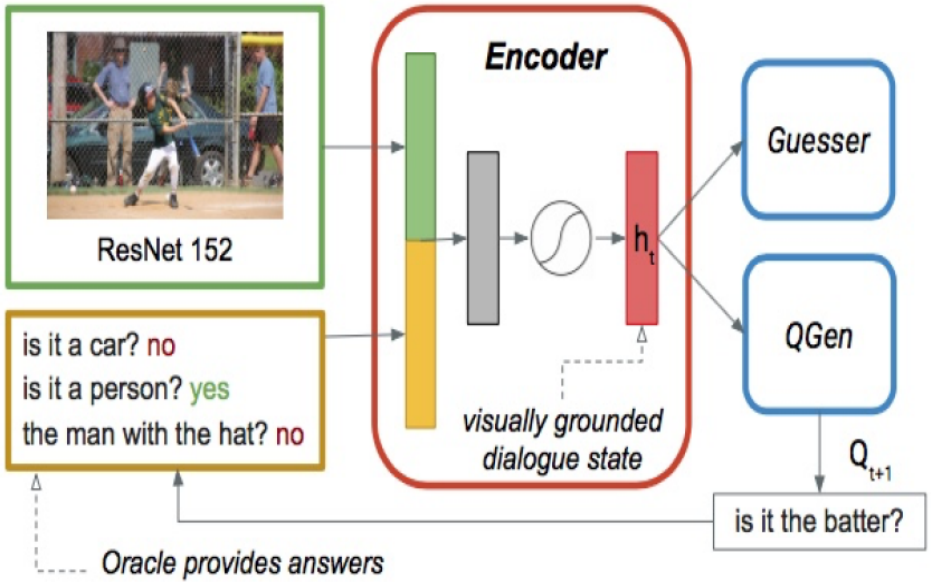


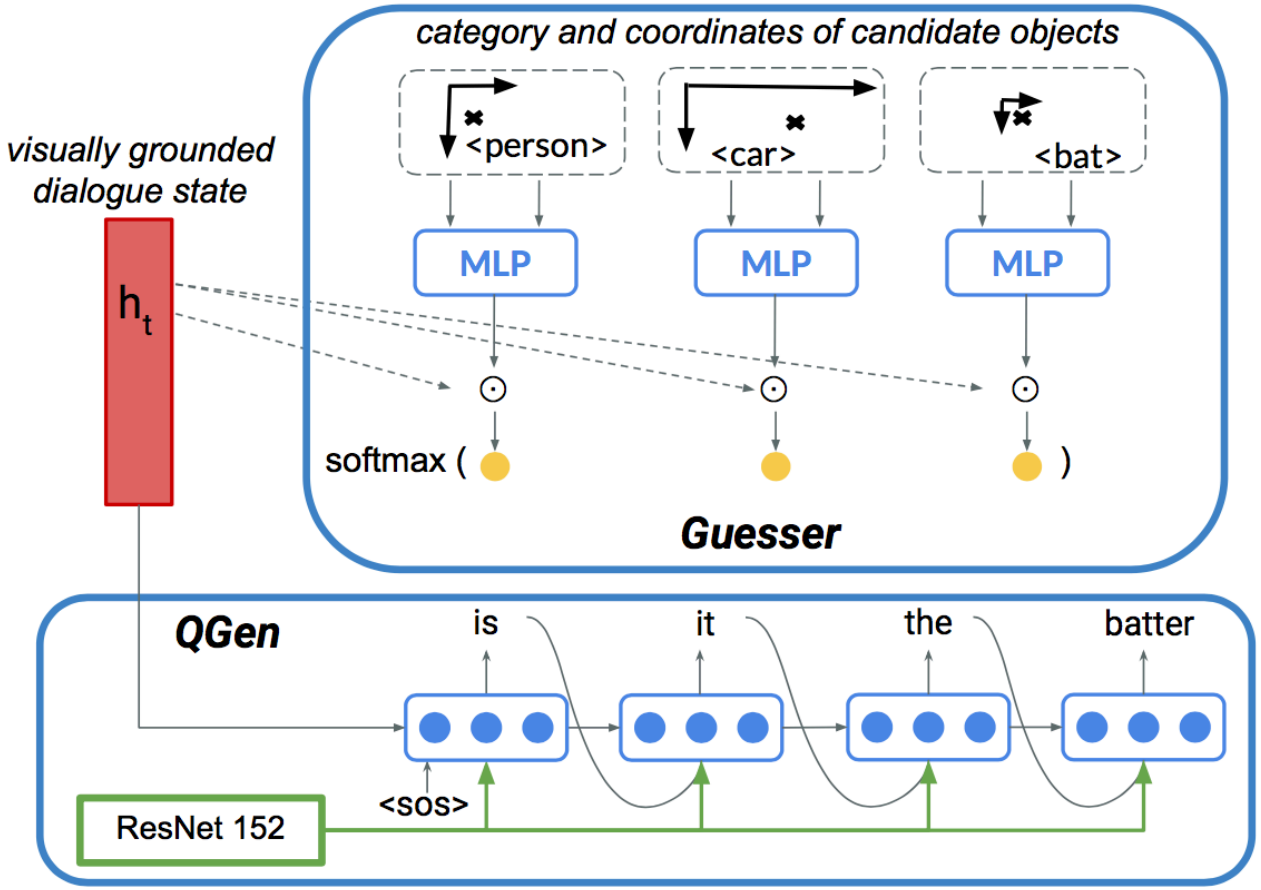
- Encoder-Decoder with Reinforcement Leaning (Strub et al. IJCAI 2017)
  - Encoder: RNN for Dialogue, VGG for Image, concatenation
  - Decoder: RNN
  - Reinforcement learning: zero-one reward



- Improvement in Reinforcement Learning
  - Word-level reward:
    - Zhao and Tresp (IJCAI 2018): Tempered Policy Gradients
      - Introduce a sampling temperature to explicitly control the strengths of the exploration.
    - Abbasnejad, et al. (CVPR 2019): Bayesian fine-grained reward
      - Bayesian reward function, parameter in upper bound to mediates exploration-exploitation
    - …
  - Question-level reward:
    - Zhang, et al. (ECCV 2018): three intermediate rewards
      - goal-achieved, progressive and informativeness
    - Shukla et al. (ACL 2019): Regularized Information Gain
      - Regularized Information Gain as reward;
    - …

- Cooperative learning (Shekhar, et al. ***NAACL2019***)
  - Encoder: LSTM for Dialogue, ResNet152 for Image, concatenation
  - Decoder: LSTM
  - Cooperative learning
    - between QGen and Guesser

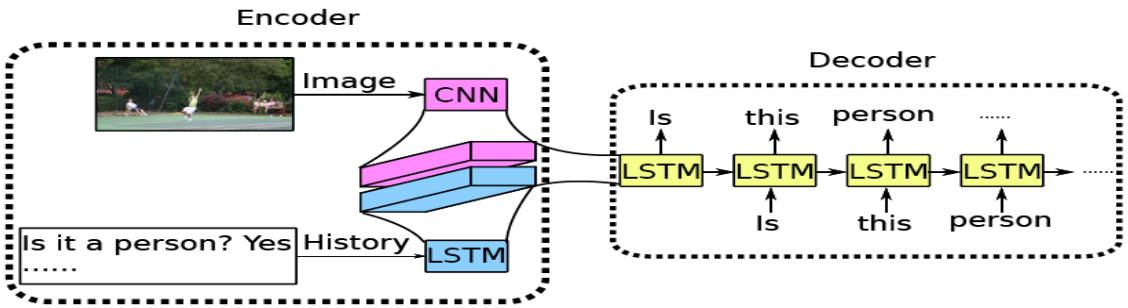




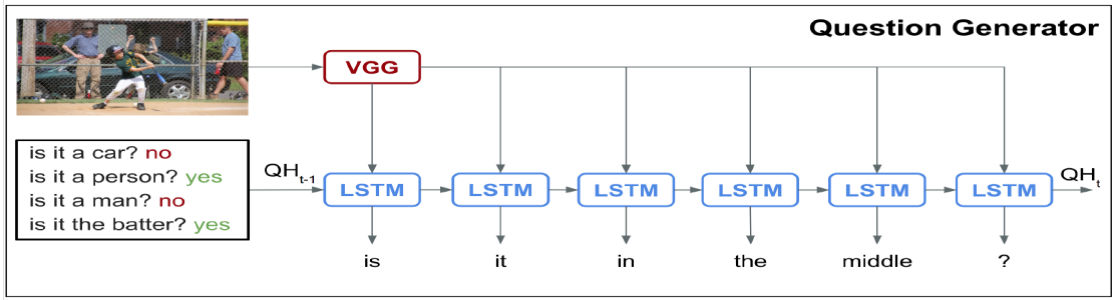
Cooperative learning

- Information based QGen
  - Lee, et al. (***NIPS2018***)：generate a question which can give maximum information gain after receiving the answer of the question (using oracle's answer)
  - Abbasnejad, (***CVPR2019***)： generate a word which has the maximum information uncertainty (using standard variance of word distribution for uncertainty)
  - Krishna, et al. (***CVPR2019***)：generate a question which maximum mutual information between question and expected answer (using category of expected oracle's answer since expected answer is unavailable)
  - Shukla, et al (***ACL2019***)：generate a question which maximum regularized information gain (using guess of guesser to measure the change of uncertainty)
  - ……

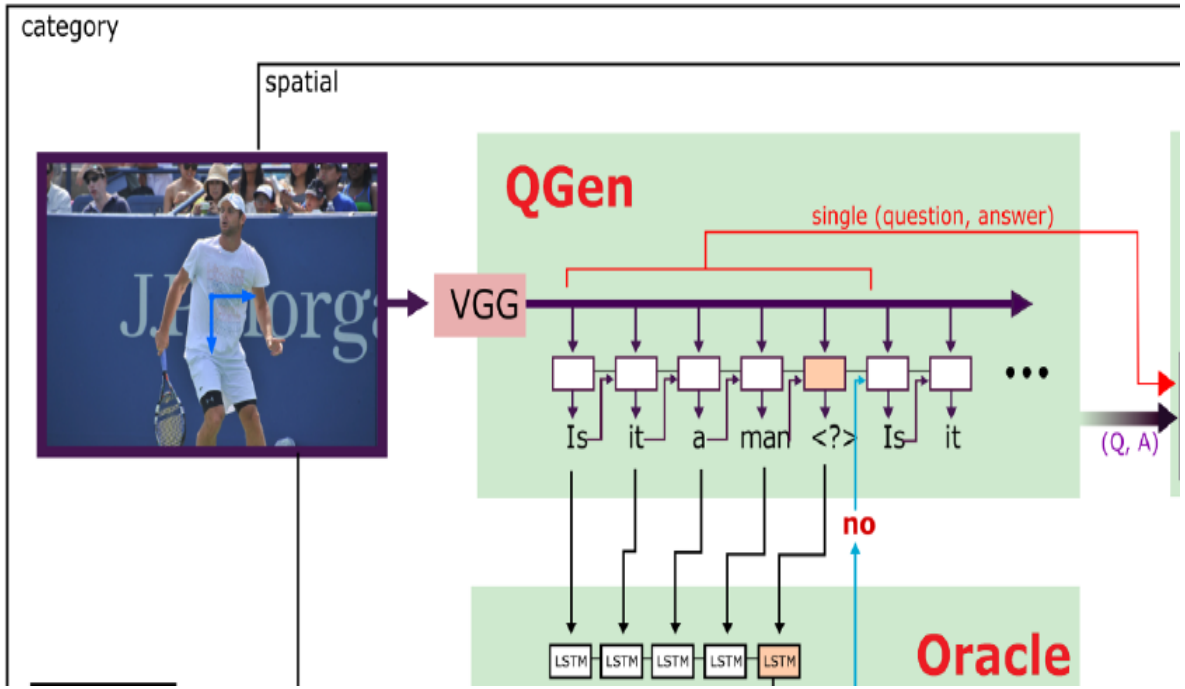
- A problem in previous work on visual representations
  - Visual representations keep unchanged through each dialogue
    - encoded images into a fixed-size vector at the beginning of a dialogue
    - used the same vector as the image feature for question generation in each round



Rui Zhao, IJCAI2018



Shekhary et al. COLING2018

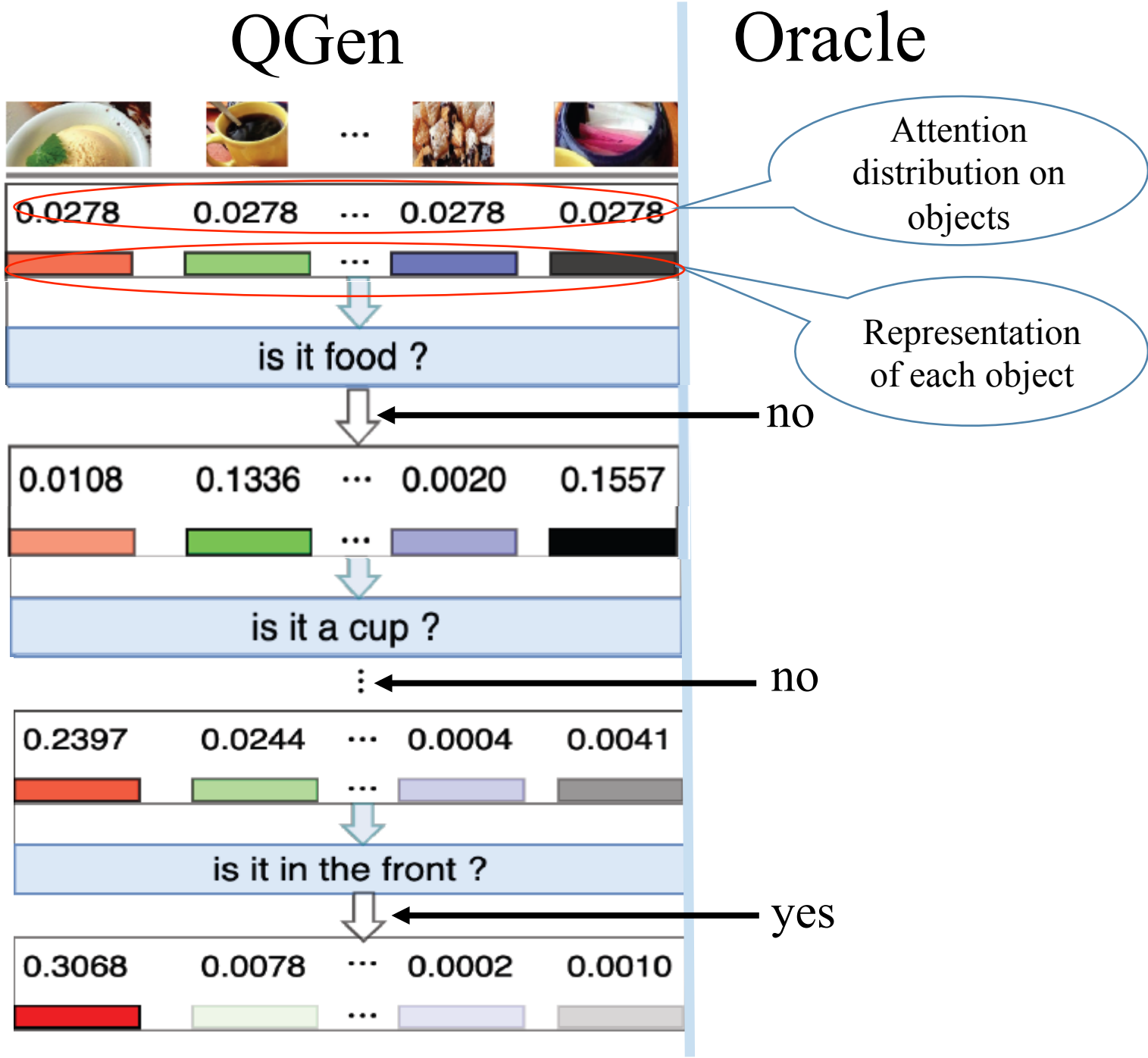


Shukla et al. ACL2019

- Motivation:
  - Attentions on different objects will change with the progress of a dialog
    - For example, When a question “Is it a person?” receives a positive answer, more attention might be paid on person objects.
  - with the change of attentions, representations of objects might be changed.
    - For example, when more attentions are paid on a objects, more detailed information of the objects might be observed. Therefore, the representation of the objects will be changed to reflect these details.

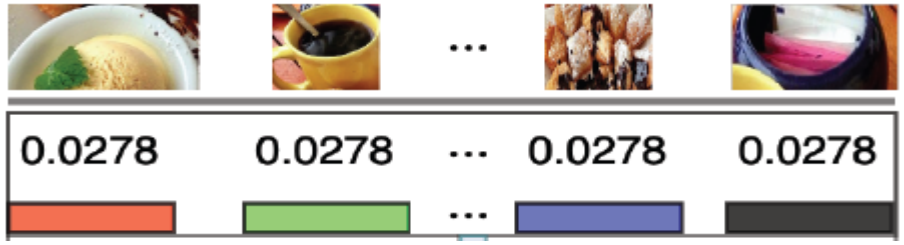
Attentions are changed: more attention might be paid on non-food objects.

representations are changed: e.g. if more attentions are paid on a object, more detailed information of the object might be observed, the representation of the objects will be changed.



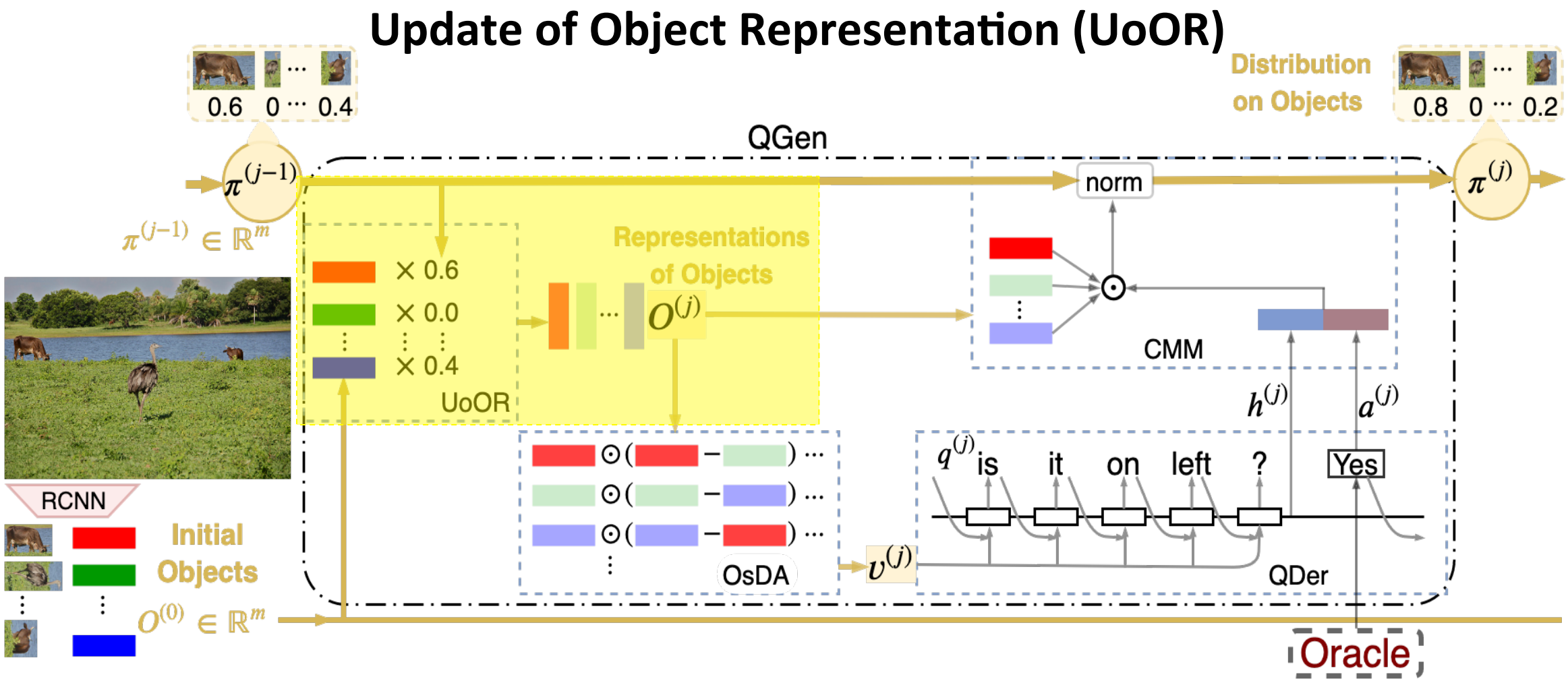
Visual Dialogue State Tracking (VDST) based QGen

- Visual Dialogue State:
  - <distribution on objects, representations of objects>

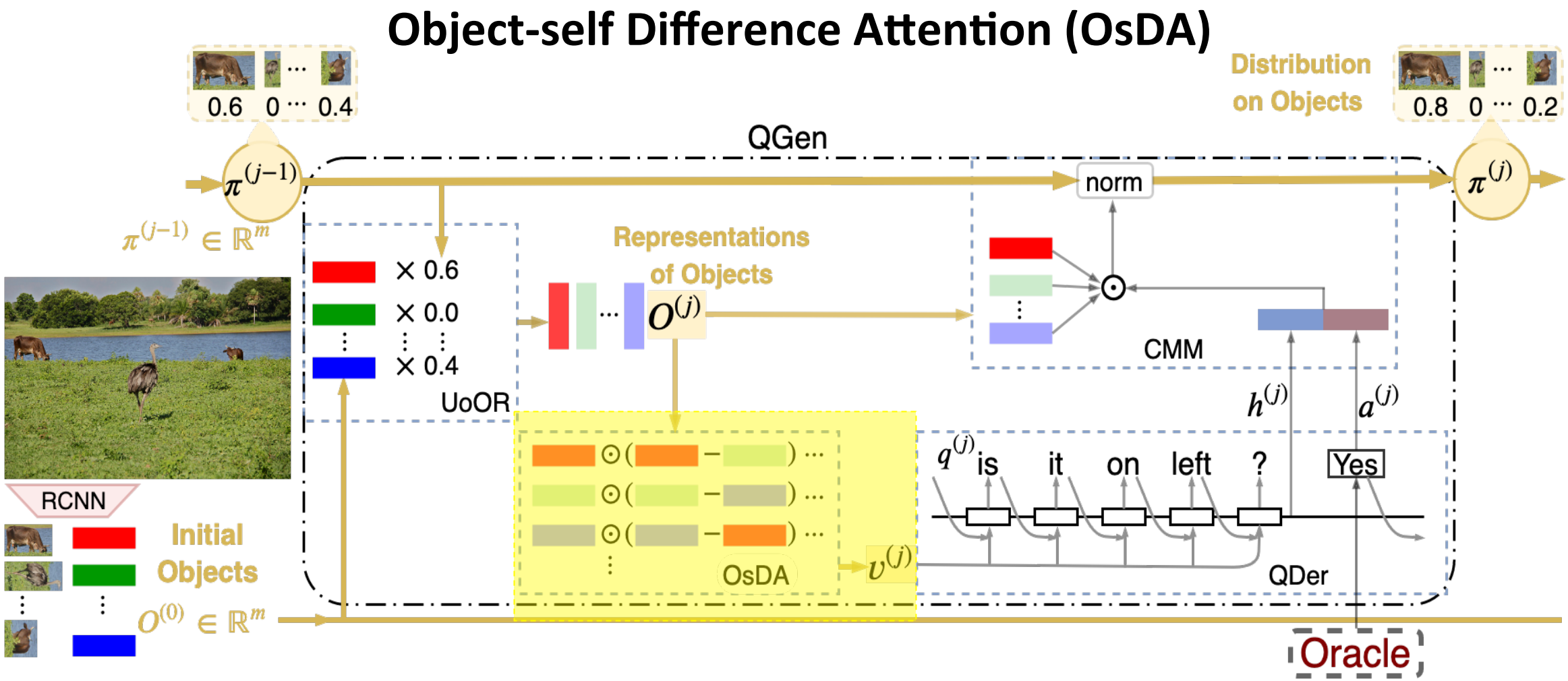


- VDST model, four modules:
  - Update of Object Representation (UoOR)
  - Object-self Difference Attention (OsDA)
  - Question Decoder (QDer)
  - Cross-Modal Matching (CMM)

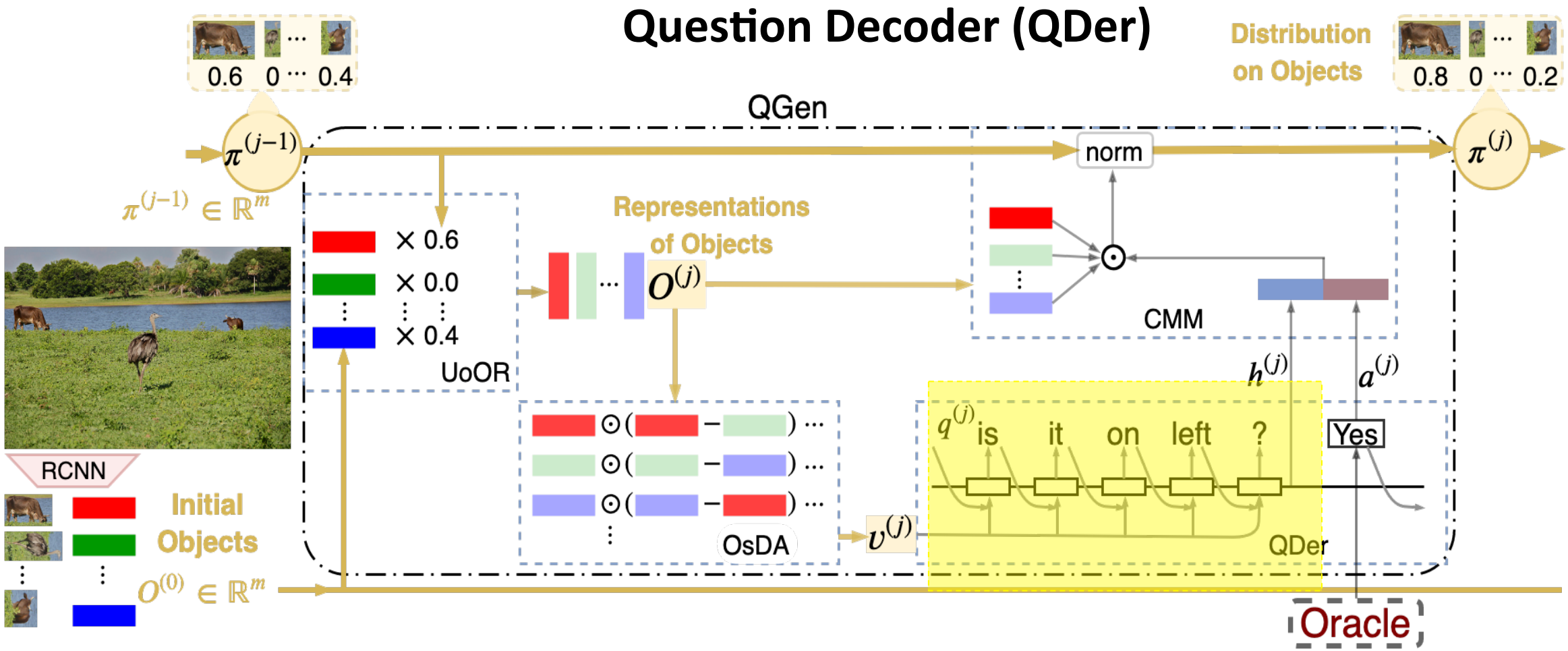
VDST based QGen



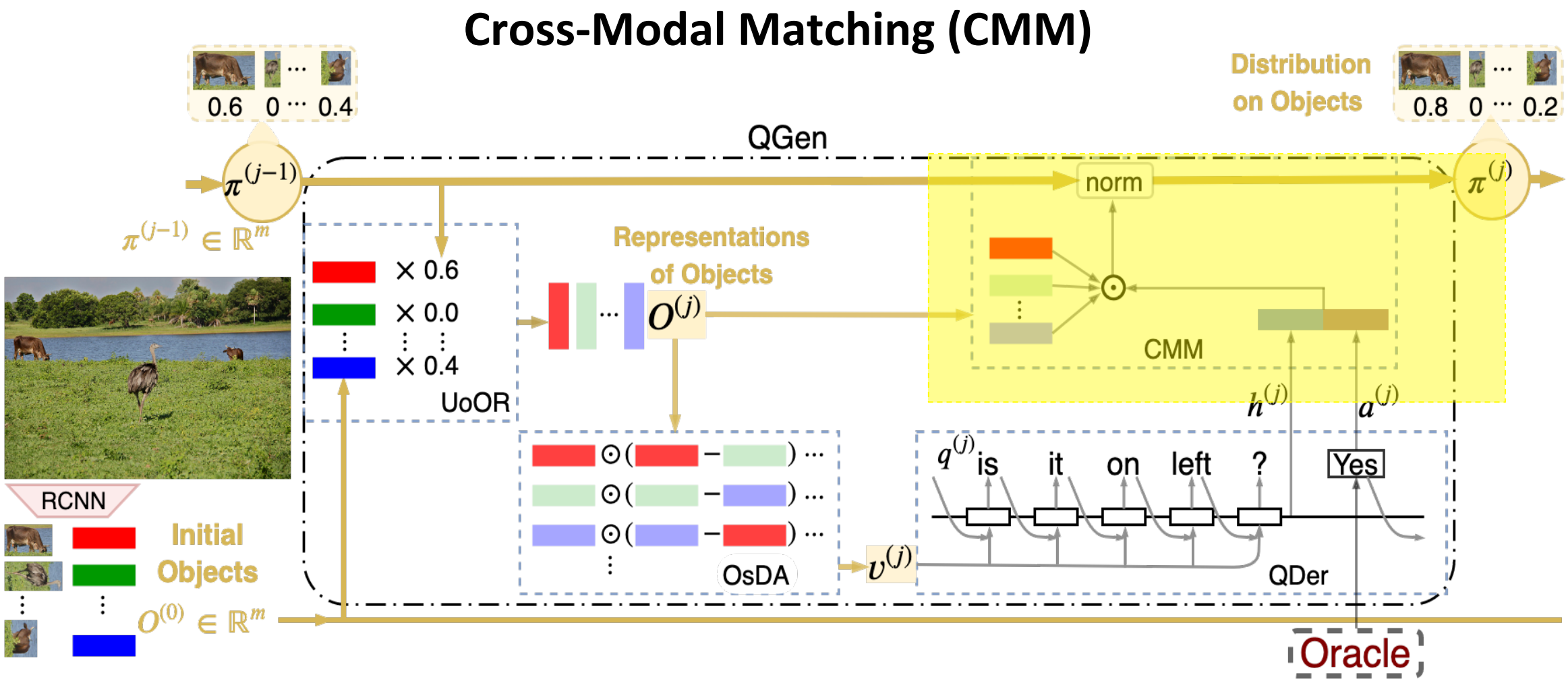
VDST based QGen



VDST based QGen



VDST based QGen



VDST based QGen

- Update of Object Representation (UoOR)

$$O^{(j)} = (\pi^{(j)})^T O^{(0)}$$

- Object-self Difference Attention (OsDA)

$$v^{(j)} = \mathrm{softmax}([o_i^{(j)} \odot (o_i^{(j)} - o_k^{(j)})]W)^T O^{(j)}$$

- Question Decoder (QDer)

$$w_{i+1}^{(j)} = \mathrm{LSTM}([v^{(j)}; w_i^{(j)}])$$

- Cross-Modal Matching (CMM)

$$\hat{\pi}^{(j)} = \mathrm{softmax}(\frac{\tanh(O^{(j)}U^T \odot V^T h_a^{(j)})}{\sqrt{d}})$$

- Updating Distribution of Objects

$$\pi^{(j)} = \mathrm{norm}(\pi^{(j-1)} \cdot \hat{\pi}^{(j)})$$

VDST based QGen

- **Training**

- Supervised learning with a cross-entropy loss

$$L_{SL}(\theta) = \sum_{t=1}^{T} -\log(p(w_t))$$

- Reinforcement learning with zero-one reward

$$L_{RL}(\theta) = -\mathbb{E}_{\pi_\theta}[\sum_{t=1}^{T}(r_A(a_t) - b)\log \pi_\theta(a_t)]$$

Experiments

- Experimental settings
  - Dataset and Evaluation
    - Same as those in de Vries, CVPR2017 & Strub et al. IJCAI 2017
  - Implementation Details
    - maximum number of questions to 5 and 8
    - maximum number of words to 12
    - …
  - Training
    - Supervised Learning: Adam optimizer
    - Reinforcement Learning: REINFORCE
  - Baseline and SOTA models
    - Listed in next table: trained with SL and RL

Experiments

- Success Rate of Guessing on New Objects

New Object				
	Sampling	Greedy	BSearch	Max Q's
SL (de Vries et al. 2017)	41.6	43.5	47.1	5
(Strub et al. 2017) (pre-trained in SL)	-	44.6	-	8
(Zhao and Tresp 2018) (pre-trained in SL)	-	48.77	-	8
VDST(ours, pre-trained in SL)	45.02	**49.49**	-	5
VDST(ours, pre-trained in SL)	**46.70**	48.01	-	8
RL (Strub et al. 2017)	58.5	60.3	60.2	5
RL (Strub et al. 2017)	62.8	58.2	53.9	8
TPG (MN)(Zhao and Tresp 2018)	62.6	-	-	5
VQG (Zhang et al. 2018)	63.2	63.6	63.9	5
ISM (Abbasnejad et al. 2018)	-	64.2	-	-
Bayesian (Abbasnejad et al. 2019)	61.4	62.1	63.6	5
RIG as rewards (Shukla et al. 2019)	65.20	63.00	63.08	8
RIG as a loss with 0-1 rewards (Shukla et al. 2019)	67.19	63.19	62.57	8
VDST(ours)	**66.22**	**67.07**	**67.81**	5
VDST(ours)	**69.51**	**70.55**	**71.03**	8

Experiments

- Success Rate of Guessing on New Games

New Game				
SL (de Vries et al. 2017)	39.2	40.8	44.6	5
SL (de Vries et al. 2017)	-	40.7	-	8
VDST(ours, pre-trained in SL)	42.92	**45.94**	-	5
VDST(ours, pre-trained in SL)	**44.24**	45.03	-	8
RL (Strub et al. 2017)	56.5	58.4	58.4	5
RL (Strub et al. 2017)	60.8	56.3	52.0	8
VQG (Zhang et al. 2018)	59.8	60.7	60.8	5
ISM (Abbasnejad et al. 2018)	-	62.1	-	-
GDSE-SL (Shekhar et al. 2019)	-	47.8	-	5
GDSE-SL (Shekhar et al. 2019)	-	49.7	-	8
GDSE-CL (Shekhar et al. 2019)	-	53.7	-	5
GDSE-CL (Shekhar et al. 2019)	-	58.4	-	8
Bayesian (Abbasnejad et al. 2019)	59.0	59.8	60.6	5
RIG as rewards (Shukla et al. 2019)	64.06	59.0	60.21	8
RIG as a loss with 0-1 rewards (Shukla et al. 2019)	65.79	61.18	59.79	8
VDST(ours)	**63.85**	**64.36**	**64.44**	5
VDST(ours)	**66.76**	**67.73**	**67.52**	8

Experiments

- Quality of the dialogues
  - Repeated Questions
  - Lexical diversity

	%Games with repeated Q's	Lexical diversity
SL (de Vries et al. 2017)	93.50	0.030
RL (Strub et al. 2017)	96.47	0.073
GDSE-SL (Shekhar et al. 2019)	55.80	0.101
GDSE-CL (Shekhar et al. 2019)	52.19	0.115
VDST(ours)	**21.9**	**0.150**
Human	-	0.731

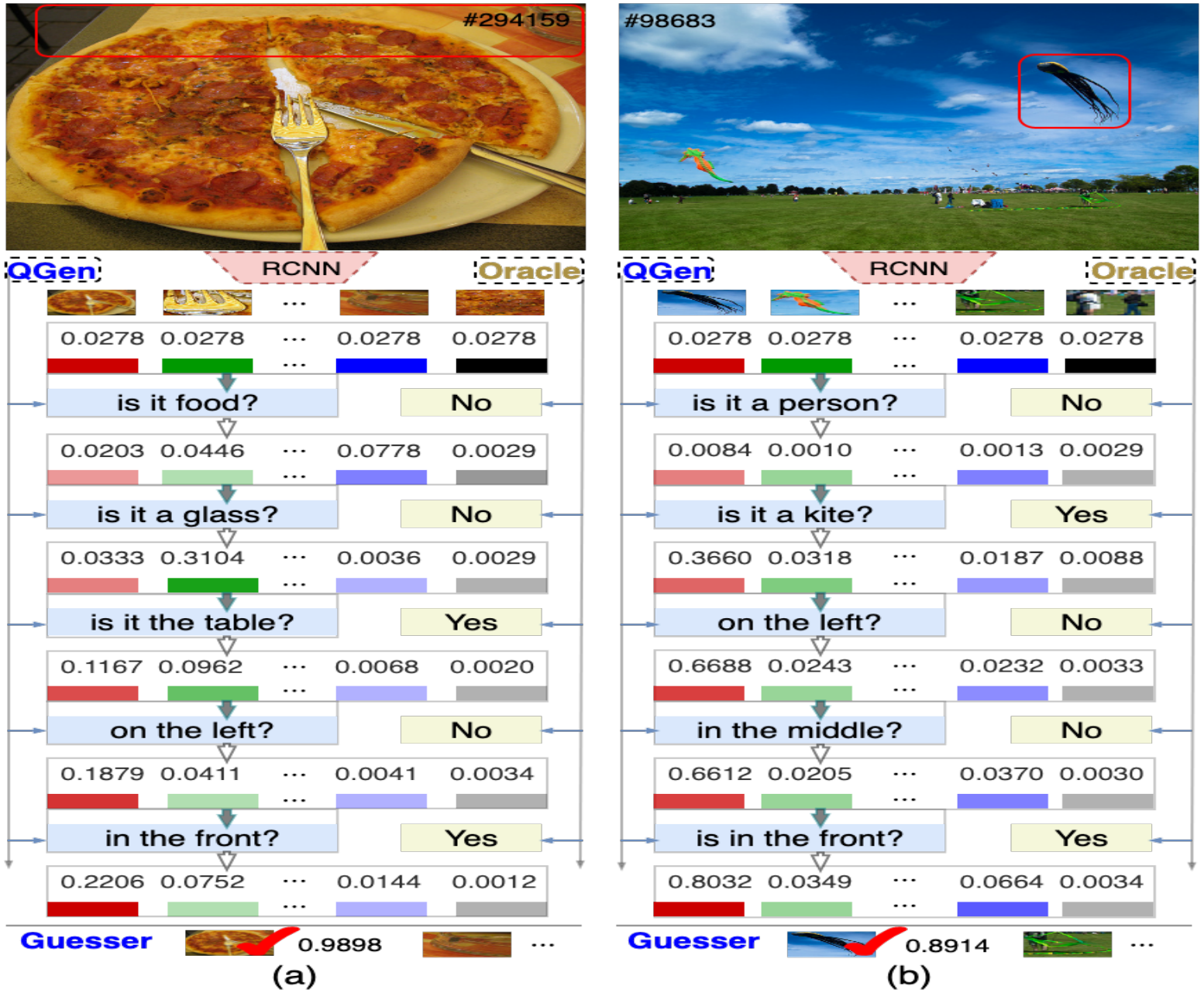
Experiments

- Ablation Study

#	Model	New Object	
		Sampling	Greedy
1	VSDT(full model)	66.22	67.07
2	−UoOR&CMM	62.73	64.30
3	−OsDA	64.06	65.21
		New Game	
1	VSDT(full model)	63.85	64.36
2	−UoOR&CMM	60.07	60.93
3	−OsDA	60.89	61.71

Experiments

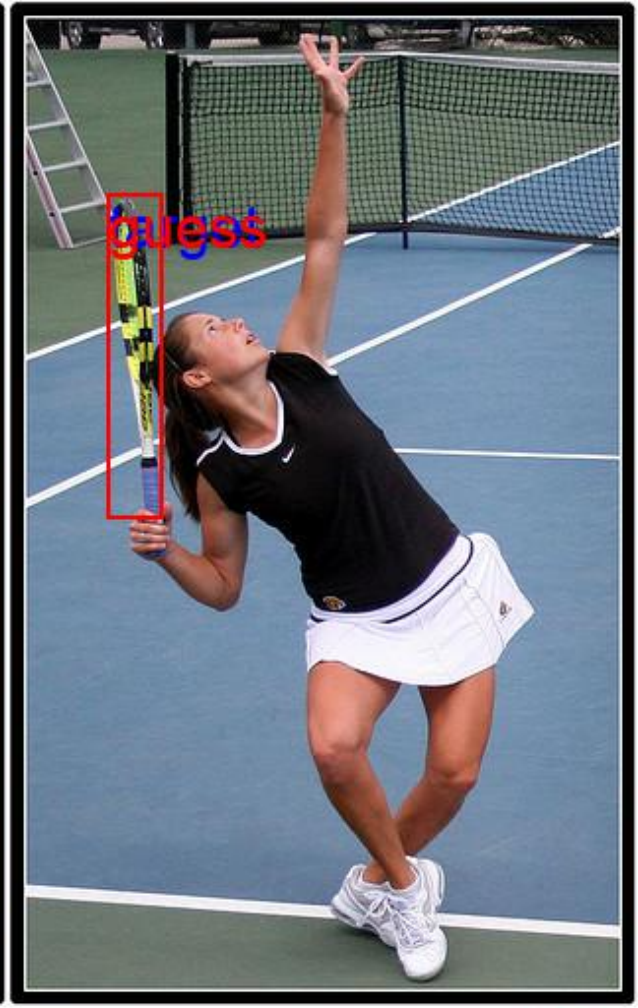
- Qualitative Evaluation
  - Two successful examples show the tracking procedure of VDST



Learned Question Policy

- Learned a simple, stable (67%) but efficient question strategy
  - ❶ Category questions, SUPER-CATE and OBJECT type
  - ❷ Keeps asking category questions until the category is decided
  - ❸ Then shift to location questions

is it a teddy ? no
is it a person ? no
is it the bed ? yes
on the right ? yes
is it in the front ? yes
Success



is it a person ? no
is it a racket ? yes
on the left ? no
in the middle ? yes
on the right ? yes
Success

is it a computer ?	no
is it a keyboard ?	no
is it a person ?	yes
on the left ?	no
is in the middle ?	no

Success

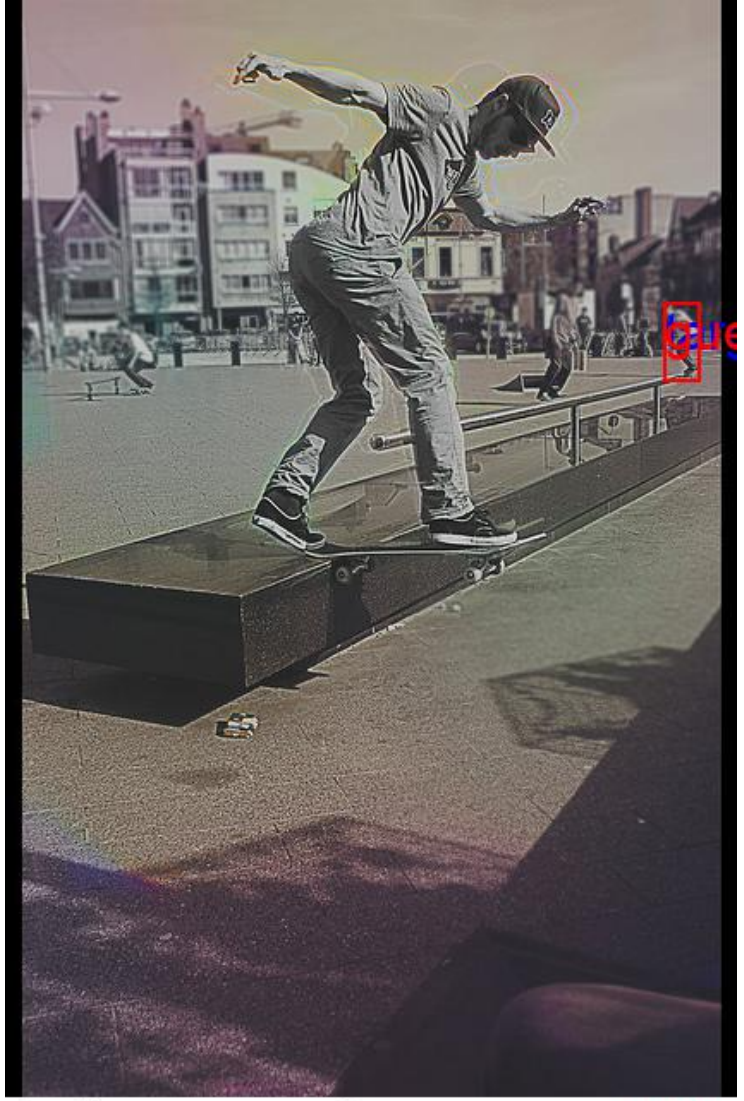
is it a cow ?	yes
is it on the left ?	no
in the middle ?	yes
is on the right ?	yes
is it in the front ?	no

Success

is it a person ?	yes
on the left ?	no
in the middle ?	no
on the right ?	yes
is it in the front ?	no

Success

is it a a phone ? no
is it a tool ? yes
on the left ? no
in the middle ? no
on the right ? yes
Success



is it a person ? yes
on the left ? no
in the middle ? no
on the right ? yes
is it in the front ? no
Success

is it a animal ? yes
is it on the left ? no
in the middle ? yes
is on the right ? no
in the middle ? yes
Success

Learned Question Policy

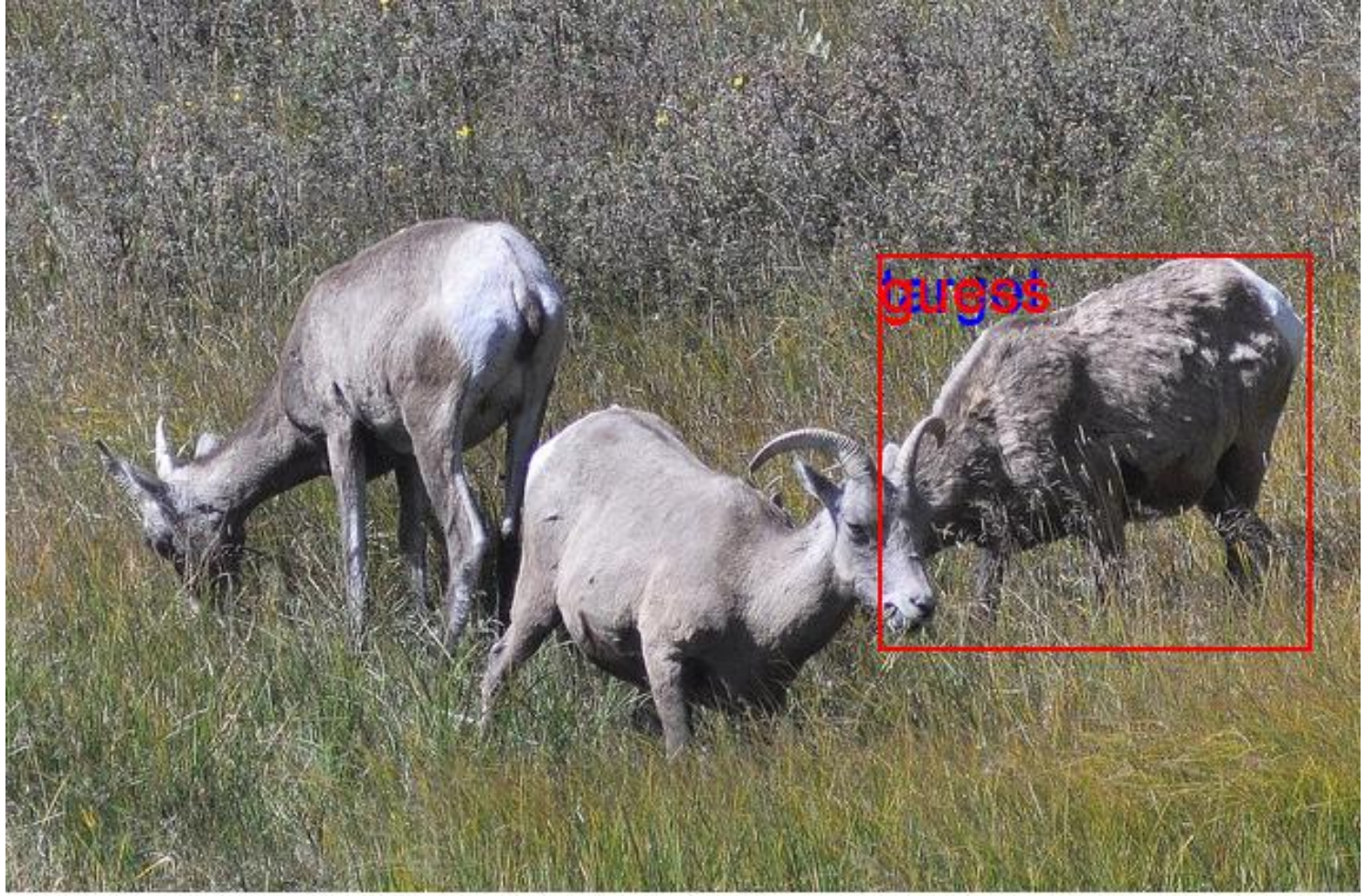
- Same questioning strategy on similar structure image

is it a person ?	yes
on the left ?	no
in the middle ?	no
on the right ?	yes
is it in the front ?	no

Success

is it a animal ?	yes
is it on the left ?	no
in the middle ?	no
on the right ?	yes
is it in the front ?	no

Success



is it a animal ?	yes
is on the left ?	no
in the middle ?	no
is on the right ?	yes
is it in the front ?	yes

Success

is it a vase ?	yes
on the left ?	no
in the middle ?	yes
on the right ?	no
in the front ?	yes

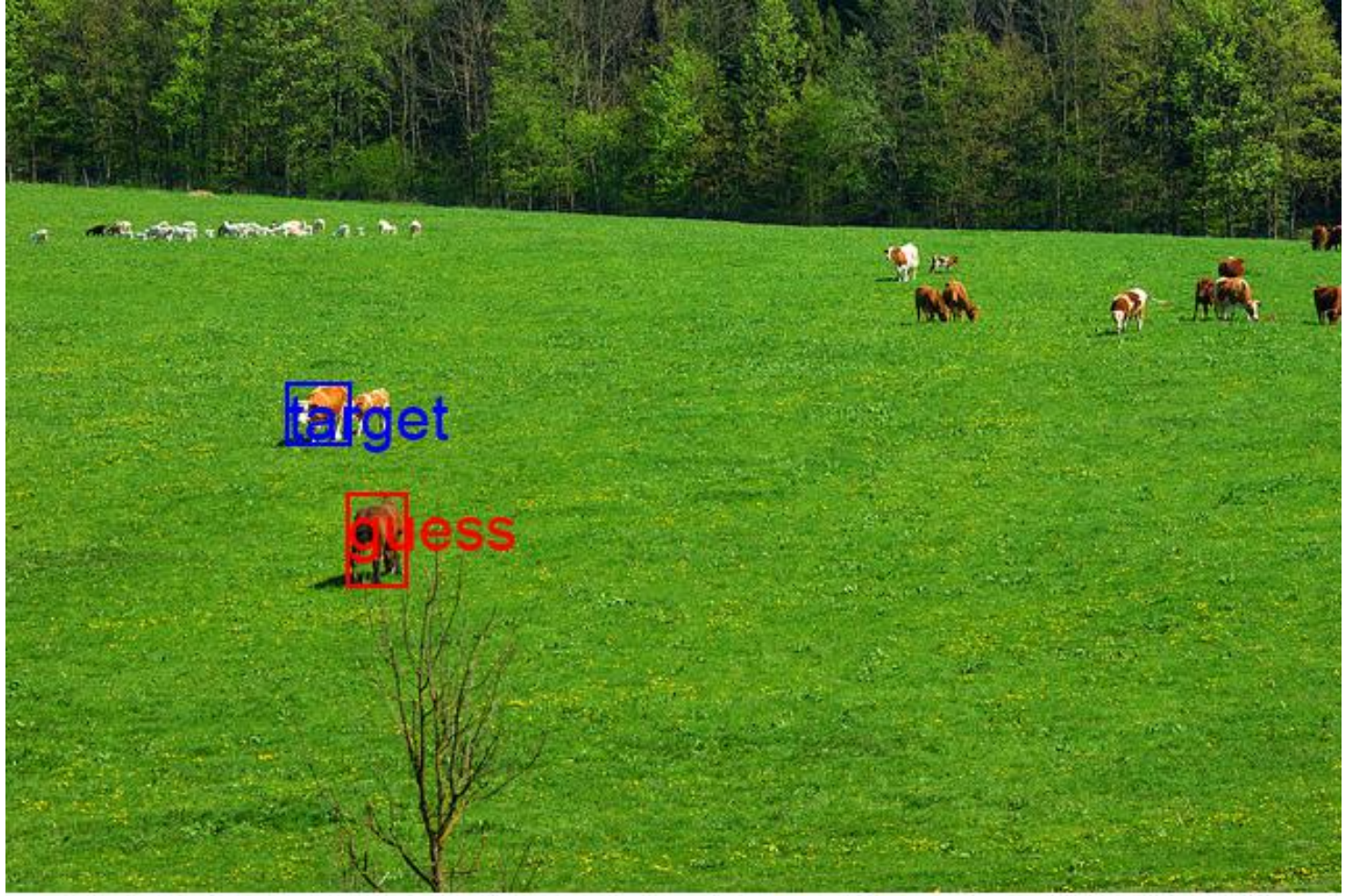
Success

Learned Question Policy

- Failure guess

is it a person ?	no
is it a board ?	yes
on the left ?	no
in the middle ?	yes
is on the right ?	yes
Failure	

is it a light ?	yes
is it on the left ?	no
in the middle ?	no
on the right ?	yes
is it in the front ?	no
Failure	



is it a cow ?	yes
is on the left ?	yes
is it in the front ?	no
on the right ?	no
in the middle ?	yes

Failure

is it a suitcase ?	yes
on the left ?	no
is it in the middle ?	yes
on the right ?	yes
the one in the front ?	yes

Failure