Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning Supplement

Abhishek Das[1,*], Satwik Kottur[2,*], José M.F. Moura[2], Stefan Lee[3], Dhruv Batra[1,4]
[1]Georgia Institute of Technology, [2]Carnegie Mellon University, [3]Virginia Tech, [4]Facebook AI Research
visualdialog.org

Abstract

This supplement is organized as follows:

- *Sec. 1 performs the following ablation study – how well would different models perform if they were provided human dialog as input? Specifically, how good are the different models on the task of human-dialog-based image retrieval?*
- *Sec. 2 describes details about hyperparameter selection and training.*
- *Finally, Sec. 3 presents qualitative results on image guessing task and generated language.*

It contains 2 figures and 2 tables:

- *Fig. 1 presents image-retrieval evaluation.*
- *Fig. 2 shows image guessing results for our model.*
- *Tab. 1 highlights example* Q-BOT-A-BOT *interactions for various models.*
- *Tab. 2 presents qualitative comparison of answers generated by various models.*

1. Dialog-based Image Retrieval

In Section 6 of the main paper, we evaluated how well models learn to cooperate at image guessing by setting up a retrieval task on the `test` split of VisDial v0.5. Specifically, we allow models to interact for 10 rounds where after each round of dialog exchange, Q-BOT predicts a feature representation $\hat{y}_t$, and we sort the entire `test` set in ascending order of distance to this prediction and compute the rank of the true image. Fig. 1a (reproduced from the main paper) shows mean percentile rank of the true image across rounds.

In this section, we conduct ablations to study how well different models perform when provided human dialog as input. Specifically, we study the performance of different models on the task of human-dialog-based image retrieval.

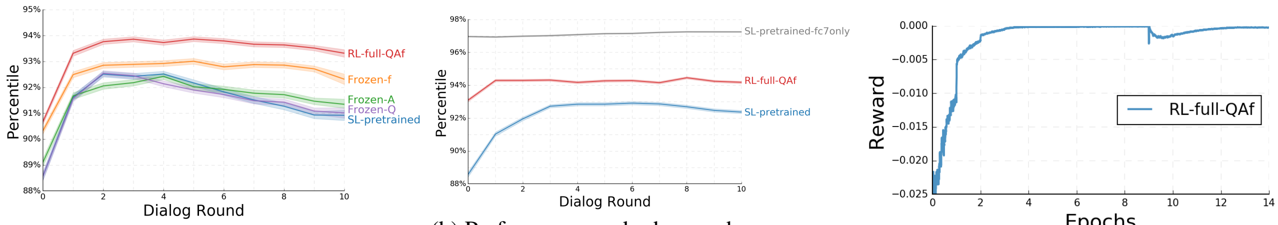
The task of human-dialog-based image retrieval allows us to probe two aspects of our models. First, we study how well RL trained models retain an understanding of human dialog from their SL-pretraining. Second, we can estimate a 'human upper bound' on the image guessing task. Specifically, when provided with a human-human dialog, A-BOT has 'perfect' perception and Q-BOT always asks consistent and non-repeating questions.

In addition to testing both models from the main paper (`SL-pretrained` and `RL-full-QAf`), we also train a supervised model, `SL-pretrained-fc7only`, specifically for this task of human-dialog-based image retrieval for comparison to those models. Specifically, a model trained to regress to image features given a human dialog. In practice, this amounts to finetuning f in `SL-pretrained` Q-BOT for image regression only, instead of a question-generation + fc7 regression multitask objective.

Fig. 1b shows results of this experiment as mean percentile rank of true image across rounds of human dialog. We note two observations. First, all models improve at the task of image retrieval when provided human-dialog, suggesting that language generation is a difficult task, and if we are provided 'good' dialog – self-consistent, accurate, and free from repetition – all models can improve at image guessing.

Second, we observe `SL-pretrained-fc7only` performs significantly better than the other models. This is unsurprising as this model is *explicitly trained* to regress from human dialog and lacks the additional utterance generation loss found in the other models. The strong performance of this model is encouraging because it implies that if we can generate better dialog, we can expect to do better at image retrieval.

[*]The first two authors (AD, SK) contributed equally.



(a) Performance under Q-BOT-A-BOT generated dialog.

(b) Performance under human-human dialog.

(c) Learning curves for `RL-full-QAf`.

Figure 1: Guessing game evaluation – **(a)** based on Q-BOT-A-BOT dialog, and **(b)** based on ground-truth human dialog. **(c)** plots reward vs. epochs for `RL-full-QAf`.

2. Implementation Details

In this section, we provide additional details about the architecture, hyperparameters, and training procedure for the policy networks described in Section 4.1.

Model Specifics. Both Q-BOT and A-BOT policies are modeled via Hierarchical Recurrent Encoder-Decoder networks [2, 5, 6] and no parameters are shared between them.

All LSTMs (fact embedding, history encoder, question/answer decoder, and question encoder) have 2 layers and 512 dimensional hidden states. Word embeddings are 200 dimensional and are shared across question, history and decoder LSTMs. All models are implemented in Torch [1].

Curriculum Training. We use the penultimate layer activations from VGG-16 [7] as image representations and models are pretrained for 15 epochs on VisDial [2]. Afterwards they are transitioned to policy-gradient training by a fixed curriculum. Supervised training is performed for the first K rounds of a dialog and policy-gradient updates are performed for the final $10-K$. We anneal K down by 1 every epoch (starting from 9). When K is zero, ground-truth captions are swapped for captions generated by an off-the-shelf captioning model [3], so that an infinite number of dialogs can be simulated without any human supervision. Models are trained for 5 epochs on these generated captions. Fig. 1c plots reward vs. epochs during policy learning. We use Adam [4] with a learning rate of 10^{-3}, and clamp gradients to $[-5, -5]$ to avoid explosion for all settings.

3. Qualitative Results

We provide additional qualitative results in Tab. 1, Tab. 2 and Fig. 2. Tab. 1 compares Q-BOT-A-BOT interactions for `SL-pretrained` and `RL-full-QAf` alongside the source images, their captions and human dialogs.

We notice that `SL-pretrained` agents often repeat questions –

"*what color are his shoes? black;*
what color are his shoes? black;
what color ..." (see first row).

`RL-full-QAf` agents do not fall into these loops as often; moreover, when they do repeat themselves, it is later in the dialog and they often sensibly recover from these loops –

"*can you see any trees? yes, there are lot of trees in background;*
can you see any buildings in background? no, I can not;
can you see any trees? ...;
does it look like they are in park?" (row 1).

We also find that `SL-pretrained` agents tend to produce generic, 'safe' responses – "*I can't tell; I don't know; Can't see*", while `RL-full-QAf` interactions are noticeably more diverse and image-discriminative – *i.e.*,

"*what are they wearing?*
they are all wearing snow pants and jackets" (row 1)

"*where is man located?*
looks like classroom of some sort" (row 9) .

Notice that the human dialog does not mention location or classroom, despite these facts being visually discriminative.

In Sec. 6 of the main paper, we quantitatively evaluate how well A-BOT mimics human dialog (using the retrieval metrics from [2]). Recall that `Frozen-Q-multi` outperforms the other approaches on VisDial answering metrics. In Tab. 2, we present qualitative examples of responses generated by `SL-pretrained`, `RL-full-QAf` and `Frozen-Q-multi` A-BOT for ground-truth human-asked questions. We again find that `SL-pretrained` responses are predictably 'safe', while `RL-full-QAf` responses are more diverse and informative; however, this sometimes leads to `RL-full-QAf` being inaccurate. `Frozen-Q-multi` strikes a balance between response specificity and human-ness –

"*how old are men?*
`SL-pretrained`: *I can't tell,*
`RL-full-QAf`: *they look like teenagers,*
`Frozen-Q-multi`: *middle-aged*" (row 2);

"*what colors are umbrella?*
`SL-pretrained`: *white,*
`RL-full-QAf`: *rainbow colors,*
`Frozen-Q-multi`: *many different colors*" (row 1).

Fig. 2 explores how Q-BOT's image guess changes through rounds of dialog. The left column shows the true image and caption and the middle column shows the text interaction between Q-BOT and A-BOT for select rounds. To show how Q-BOT's prediction changes through the dialog we show the five nearest neighbors of the prediction in the right column, with the nearest neighbor outlined in red. In the middle column we show where this nearest neighbor image ranks among other images in distance to the true feature, also showing other similarly distant images for context.

References

[1] Torch. http://torch.ch/. 2

[2] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual Dialog. In *CVPR*, 2017. 2, 4, 5

[3] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[4] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 2

[5] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 2016. 2

[6] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. *arXiv preprint arXiv:1605.06069*, 2016. 2

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

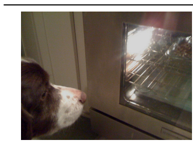
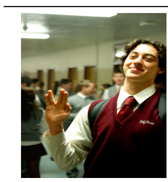
Image + Caption	Human-Human dialog [2]		SL-pretrained Q-BOT-A-BOT dialog		RL-full-QAf Q-BOT-A-BOT dialog	
	What are their genders?	1 man 1 woman	How many people?	2	Are people male or female or both?	I think 1 is male, can't see other 1 ...
	Are they both adults?	Yes	Male or female?	I can't tell	What are they wearing?	They are all wearing snow pants and jackets
	Do they wear goggles?	Looks like sunglasses	What are they wearing?	I can't tell, they are far away	Is it snowing?	It does not appear to be snowing at moment
	Do they have hats on?	Man does	What color are skis?	I can't tell	Can you see any trees?	Yes, there are lot of trees in background
	Are there any other people?	No	Are they wearing goggles?	Yes	Can you see any buildings in background?	No, I can not
	What color is man's hat?	Black	Are they wearing goggles?	Yes	Can you see any trees?	Yes, there are lot of trees in background
	Is it snowing now?	No	Are they wearing goggles?	Yes	Can you see any buildings in background?	No, I do not see any buildings
	What is woman wearing?	Blue jacket and black pants	Are they wearing goggles?	Yes	Can you see any trees?	Yes, there are lot of trees in background
	Are they smiling?	Yes	Are they wearing hats?	Yes	Can you see any buildings in background?	No , I do not see any buildings
A couple of people in the snow on skis.	Do you see trees?	Yes	Are they wearing goggles?	Yes	Does it look like they are in park?	It does not appear to be
	How old are men?	They look to be 30's	How old are men?	They look like they are in their 20's	How old are men?	They look like they are in their 20's
	Are they in restaurant?	I can't tell, its close up	Are they male or female?	1 male and 1 female	Are they wearing chef's hat?	No
	Do men look they might be related?	Maybe	What are they wearing?	T shirt and shorts	Do they have long hair?	No
	Are they each holding slice of pizza in 1 hand and giving thumbs up with other?	1 holding slice, other holding box...	What kind of pizza is it?	Looks like pepperoni	What kind of pizza is it?	Looks like pepperoni
	What kind of pizza do they have?	cheese	Does it look good?	Yes it does	Does it look good?	Yes
	Does box have pizza company name on it?	No	Can you tell what kind of pizza it is?	Looks like pepperoni	Are there any other people around?	No
	Are there any pictures or graphics on box?	No	Are there any toppings on pizza?	Pepperoni	Are they inside or outside?	Inside
	Are men slender or heavy set?	Slender	Are there any toppings on pizza?	Pepperoni	Can you tell what time of day it is?	No
	Do either of men wear glasses?	Yes	Are there any toppings on pizza?	Looks like cheese, but I can't tell...	Are there any other people around?	No
Two men giving thumbs up holding large slices of pizza.	1 or both?	1	Are there any other toppings on pizza?	No	What color are walls?	White
	How old is man?	In his teens	How old is man?	I can't see his face, I don't know	What color is skateboard?	It is hard to tell but I think it's black
	Is he wearing helmet?	No	What is he wearing?	shorts and t shirt	Is he wearing helmet?	Yes
	Is he wearing any protective gear?	No	What color is his shirt?	white	Is he wearing knee pads?	No, no knee pads
	What color is his hair?	Dirty blonde or dark brown	what color are his pants?	black	Is it day or night?	It is daytime
	What color is his shirt?	Black with red letters	what color is his shirt?	blue	Is it sunny?	It is bright out, but I see no shadows, so I can not
	Is he wearing shorts?	Yes	What color are his shoes?	Black	Can you see any other people?	There are 2 other people in background
	What color are his shorts?	Denim	What color are his shoes?	Black	Is it day or night?	It is daytime
	Are there other people visible?	No	What color are his shoes?	Black	Is it sunny?	Yes, it is sunny
	What color are his shoes?	I can't see them	What color are his shoes?	Black	What color is skateboard?	It is hard to tell but I think it's black
A man riding a skateboard down the side of a ramp.	Is this at skate park?	Yes	What color are his shoes?	Black	Is he wearing helmet?	Yes, he is wearing helmet
	What color is plane?	White and blue	What color is plane?	White	What color is airplane?	White with red and blue stripes
	Is there writing on plane?	Yes it says lufthansa	Is it large or small?	Large	Can you tell what airline it is?	No, it's too far away
	What color is ground?	Grey tarmac	Any people?	No	What time of day is it?	It is daytime
	Any people?	No people visible	Any writing on plane?	Yes	Can you see any people?	No
	Can you see airport?	Yes	Any writing on plane?	Yes	Can you see any buildings?	Yes, in background
	What color is it?	Grey	Any writing on plane?	Yes	Is it day or night time?	It is daytime
	Is there sunshine?	No it's kind of overcast	Any writing on plane?	Yes	Can you see sky?	Yes
	Any other planes?	No other planes	What does it say?	UNK	Are there clouds in sky?	Yes, few
	Any grass?	No grass	What color is plane?	White	Does it look like it's going to rain?	Yes it does
An airplane sitting on an asphalt landing strip.	Any trees?	No tree	Any writing on it?	Yes	Does it look like it's going to rain?	Yes it does
	What color is dog?	Brown and white	What kind of dog is it?	I can't tell	What kind of dog is it?	Terrier of some kind
	What breed is dog?	Look like brittany spaniel	What color is it?	White	What color is it?	White and grey
	Is dog wearing collar?	Not able to see	Is it large or small?	Large	Can you tell what kind of food it is?	No
	Can you tell what is cooking?	Look like beans	What kind of food is it?	I can't tell	Are there people around?	No people
	Is oven stainless steel?	Yes	Is it large or small?	Small	Is it inside or outside?	Inside
	Does it have see through glass?	Yes	What color is it?	White	Can you tell what time of day it is?	No I can't
	How many racks in oven?	2	Is it large or small?	Large	What color is room?	I can't see walls
	Can you see knobs?	No	What color are walls?	White	Do you see any windows?	No
	Does oven have wooden handle?	Not able to see	What kind of floor?	Can't see floor	Do you see any windows?	No
A dog watching food cook in an oven.	Are beans in pot?	Glass dish	What kind of floor?	Can't see floor	Do you see any windows?	No
	Are there more people in this photo or just man?	No	How old is man?	I can't see his face, I don't know	How old is man?	He looks to be in his 20's
	Just tennis player man in this photo?	Yes	What color is his hair?	Brown	What color is his hair?	His hair is brown
	Is it outdoor court?	Can't tell, it's close up	What is he wearing?	Shorts and t shirt	Is he wearing glasses?	No he is not
	What is man wearing?	White shirt and shorts	What color is his shirt?	White	Is he wearing hat?	No he is not
	Can you see net?	No	What color is his shirt?	White	Is he wearing glasses?	No he is not
	Can you see bottom of court?	Yes	What color is his shirt?	White	Are there other people around?	No it's just him
	Can you see ball in photo?	Yes	What color is his shirt?	White	Can you see net?	No I can't
	Is it close to man's racquet?	Yes	What color is his shirt?	White	What color is court?	Court is green
	Is man about to take swing?	Yes	What color is his shirt?	White	Is it sunny?	Yes it is
A man prepares to swing at the tennis ball.	Is he left handed or right handed?	Right	What color is his shirt?	White	Do you see any trees?	No I do not
	Is giraffe adult?	I cannot tell, I only see face	Is there only 1 giraffe?	Yes	Is there only 1 giraffe?	Yes
	Is grass at higher altitude than its feet?	No	Are there any other animals?	No	Are there any other animals?	No
	Is rock big?	Yes	Are there any people?	No	Are there any people?	No
	How is weather?	Sunny	Are there any other animals?	No	Does this look like zoo?	Can't tell
	What color is rock?	Tan color	Is it day or night?	Day	Is there fence?	No
	Is grass green or dry?	Green but looks little dry	Is it sunny?	Yes	Is it day or night?	Day
	Does giraffe look healthy?	Yes	What time of day it is?	I can't tell	Is it sunny?	Yes
	Does giraffe look happy?	It looks tired	Are there any other animals?	No	What time of day is it?	Morning
	Is giraffe laying down?	Yes	Are there any trees?	No	Are there any trees?	No
A giraffe is resting its head in the grass near a rock.	Is giraffe's mouth closed?	Yes	Are there any other animals?	No	Can you see sky?	No
	Are there any people in picture?	No, there are no people	Are there any people?	No	Is plate large or small?	Large
	Is cheese white?	Yes, cheese is white	What color is plate?	White	Is it on table?	I can't tell
	Are crackers white?	No, cracker is brownish	Is plate on table?	I can't tell	What kind of donut?	Plain with sugar on it
	Are there lots of greens?	No, only 1 sprig	What color is plate?	White	What kind of chips?	Potato chips
	Was picture taken in kitchen?	I cannot tell, it is too zoomed in	Is plate on table?	I can't tell	Any people?	No
	How many crackers are there?	There is 1	Is there anything else on table?	No	Is this inside or outside?	I can't tell, it's close up
	Is there design on plate?	No, plate is plane white	Is there anything else on table?	No	What color is table?	White
	Is there good lighting?	Yes, lighting is decent	Is there anything else on table?	No	Any chairs?	No
	Is plate on table?	I cannot tell I can only see plate as background	Is there anything else on table?	No	What color is table?	I can't tell
A cracker with shredded cheese and leafy greens on a white plate.	Does cracker look delicious?	Absolutely	Is there anything else on table?	No	Is it large or small?	I can't see plate
	How old is this man?	He looks like late teens	How old is man?	He looks like he's in his 20s	Where is man located?	Looks like classroom of some sort
	Does he have short hair or long hair?	Medium	What is he wearing?	T shirt and jeans	What is man wearing?	Black t shirt and jeans
	What color is his hair?	Brown	What color is his hair?	Black	Are there any other people in picture?	Yes, quite few people in background
	Is he wearing glasses?	No	What is he wearing?	T shirt and jeans	What color is table?	Picture is black and white, but it's wood table
	Does he have facial hair?	No	What color is his hair?	Black	Is there anything else on table?	Not that I can see
	How is he dressed?	He's white dress shirt with red tie and matching red vest over it looks like school uniform	What is he wearing?	T shirt and jeans	What are other people doing?	They are standing in front of him, but I don't know what they
	Is he larger man or is he skinny?	He's skinny	What color is his hair?	Black	Are there any windows?	Not that I can see
	Is this picture taken outside?	Inside	Is it short or long?	Short	What color are walls?	I can't see walls
	Can you tell if it's daytime?	I'm not sure	What color is his hair?	Black	What time of day do you think it is?	I can't tell
A man making the live long and prosper sign from star trek.	Can you tell what room he is standing in?	He's in hallway	Is it short or long?	Short	What is man wearing?	Black t shirt and jeans

Table 1: Selected examples of Q-BOT-A-BOT interactions for SL-pretrained and RL-full-QAf. RL-full-QAf interactions are diverse, less prone to repetitive and safe exchanges ("can't tell", "don't know" *etc.*), and more image-discriminative.

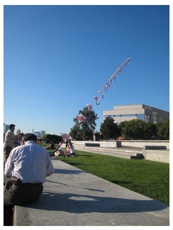
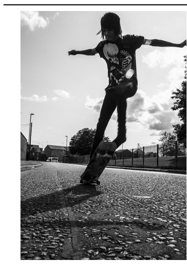
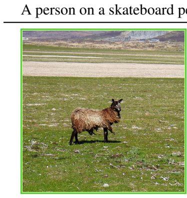
Image + Caption	Human question [2]	Human answer [2]	`SL-pretrained`	`RL-full-QAf`	`Frozen-Q-multi`
	**Round 1**: What colors are umbrella?	It is blue, red, yellow and green	White	Rainbow colors	Many different colors
	Round 2: What kind of tower is it?	I can't really tell	Not sure	It's small building	Small 1
	Round 4: Can you see any clouds?	Very few	I can not see sky	Sky is not visible	No
	Round 9: Is there any writing on tower?	There isn't any writing it's just gray	No	Not that I can see	No, there is no writing
Multi-color umbrella floating in the air next to grey tower.					
	**Round 2**: How many people?	I see 3	Too many to count	2	3
	Round 5: What are they wearing?	Men are in dress pants and shirt and woman in jeans tank	I can't tell	Jackets	Uniforms
	Round 8: How old are men?	30	I can't tell	They look like teenagers	Middle aged
People sitting outside along a concrete wall on a sunny day.					
	**Round 2**: What color is cat?	He is mostly cream colored with some tan markings	Black	White	Black and white
	Round 4: Are there any appliances?	Just microwave that I can see	Stove	Stove, stove	No
	Round 9: What else is there?	Knife block and some recipe binders	Nothing else	Toilet, toilet, toilet and top of toilet paper	Wall
A kitchen scene with a cat sitting on the counter in front of a microwave.					
	**Round 1**: How old is person?	I think around 17	Can't tell	Teenager	20s
	Round 6: Where do you think he is?	It looks like neighborhood road	I don't know	At park	In park
	Round 10: What else can you tell me about this image?	He's very skinny boy	There is snow on ground	There are lot of people in background	I see trees and sky
A person on a skateboard performing a trick.					
	**Round 5**: Are there any people?	There aren't any people there	There are no people	No people	No
	Round 8: Does lamb appear to be healthy?	Lamb looks to be very healthy	Yes, it does	Yes it does	Yes
	Round 9: What season does it seem to be?	It looks to be spring	Summer	It looks like summer	Spring
A small brown lamb walking across a lush green field.					

Table 2: Selected examples of answers generated by `SL-pretrained`, `RL-full-QAf` and `Frozen-Q-multi` A-BOT for ground-truth human-asked questions. Although differences are subtle, `SL-pretrained` answers are usually 'safe' and repetitive, `RL-full-QAf` answers diverse and sometimes unfaithful, and `Frozen-Q-multi` answers hit a balance in between.

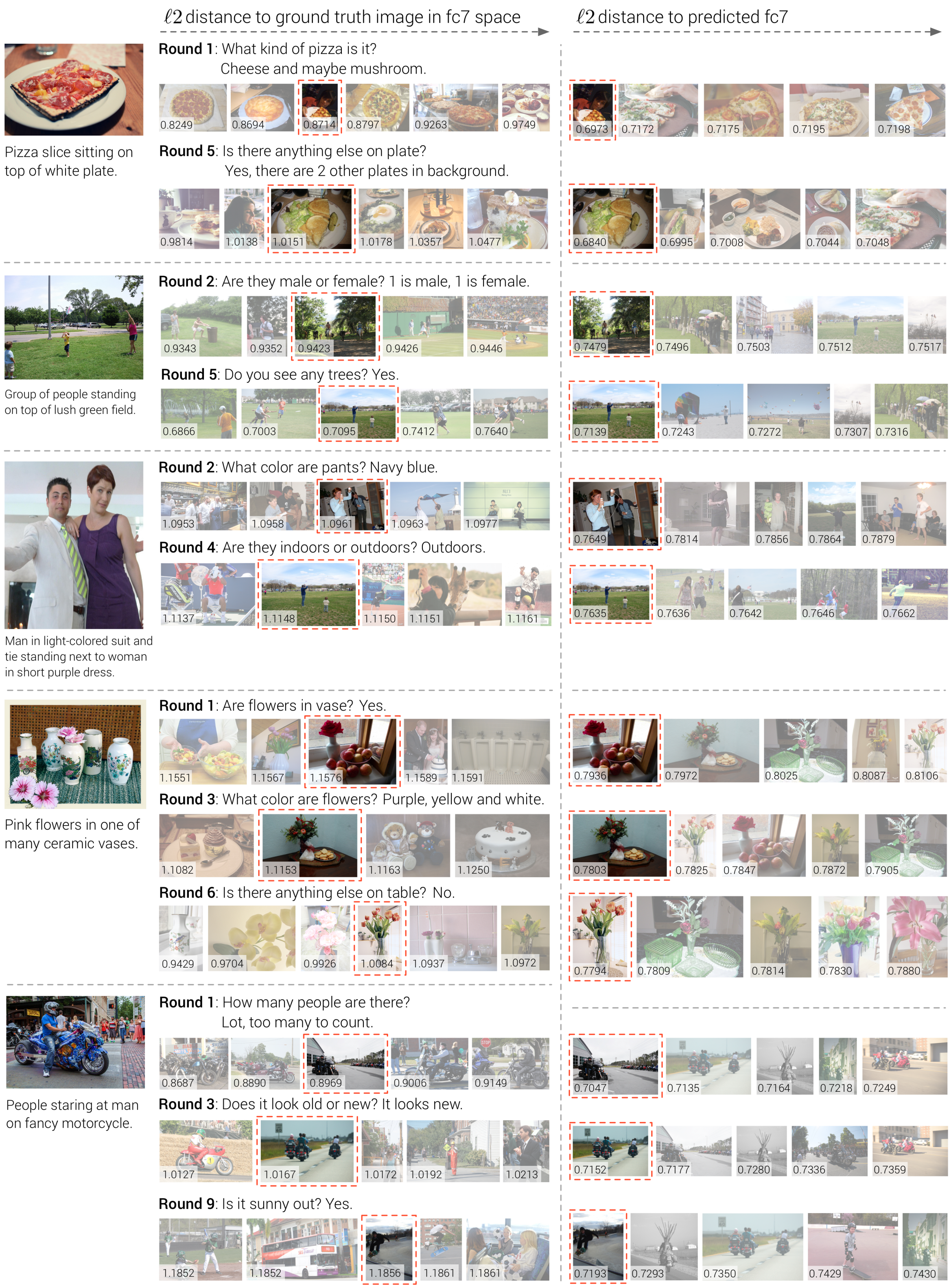


Figure 2: Qualitative results on predicted fc7-based image retrieval. Left column shows true image and caption, middle column shows dialog exchange, and a list of images sorted by their distance to the ground-truth image, right column shows a list of images sorted by their distance to predicted fc7. The image predicted by Q-BOT is highlighted in red. We can see that the predicted image is often semantically quite similar.