

A. User interface

Figure 8, 9 presents the instructions for the oracle and questioner before they started their first game.

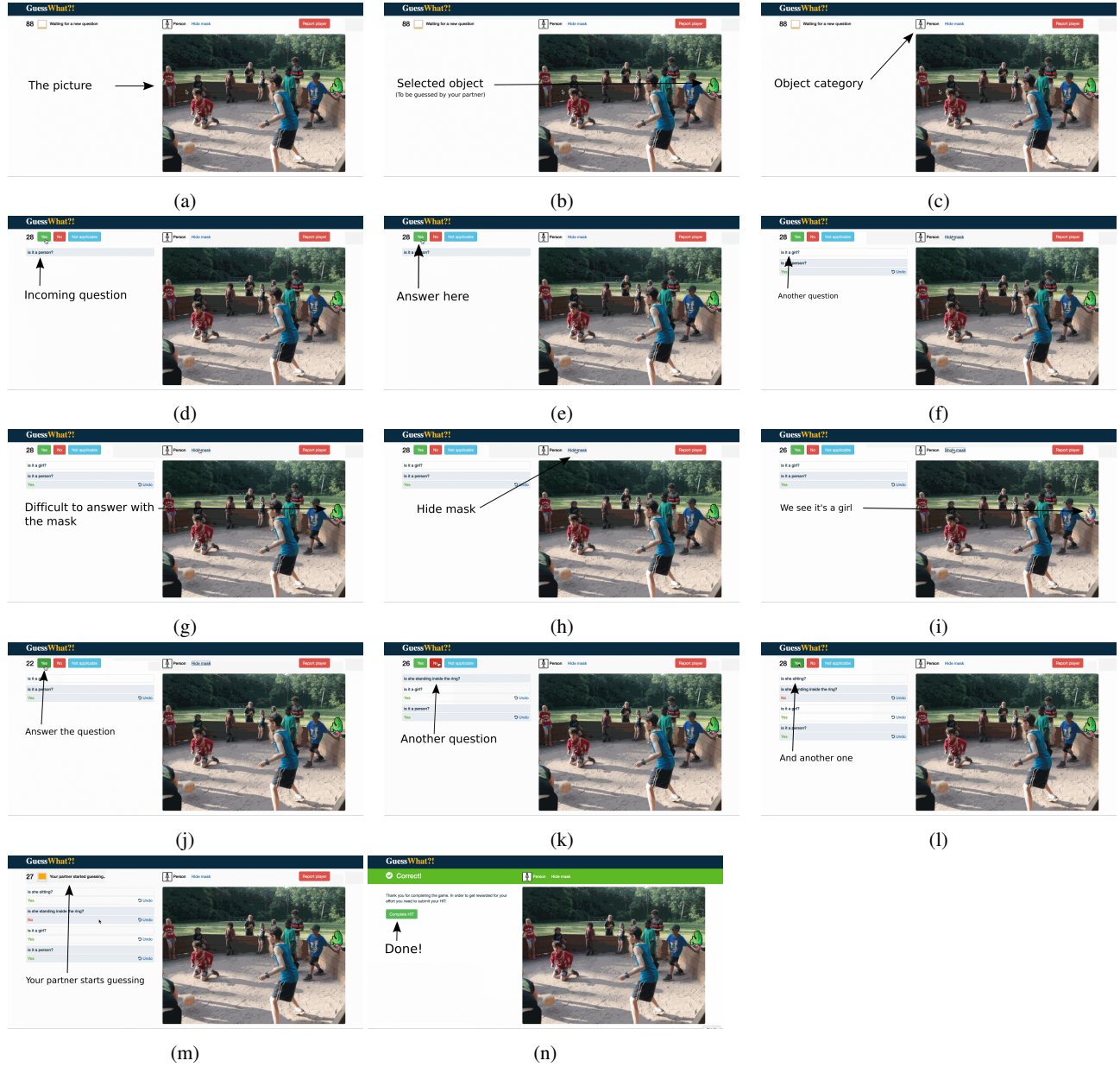


Figure 8: An example game from the perspective of the oracle. Shown from left to right and top to bottom.

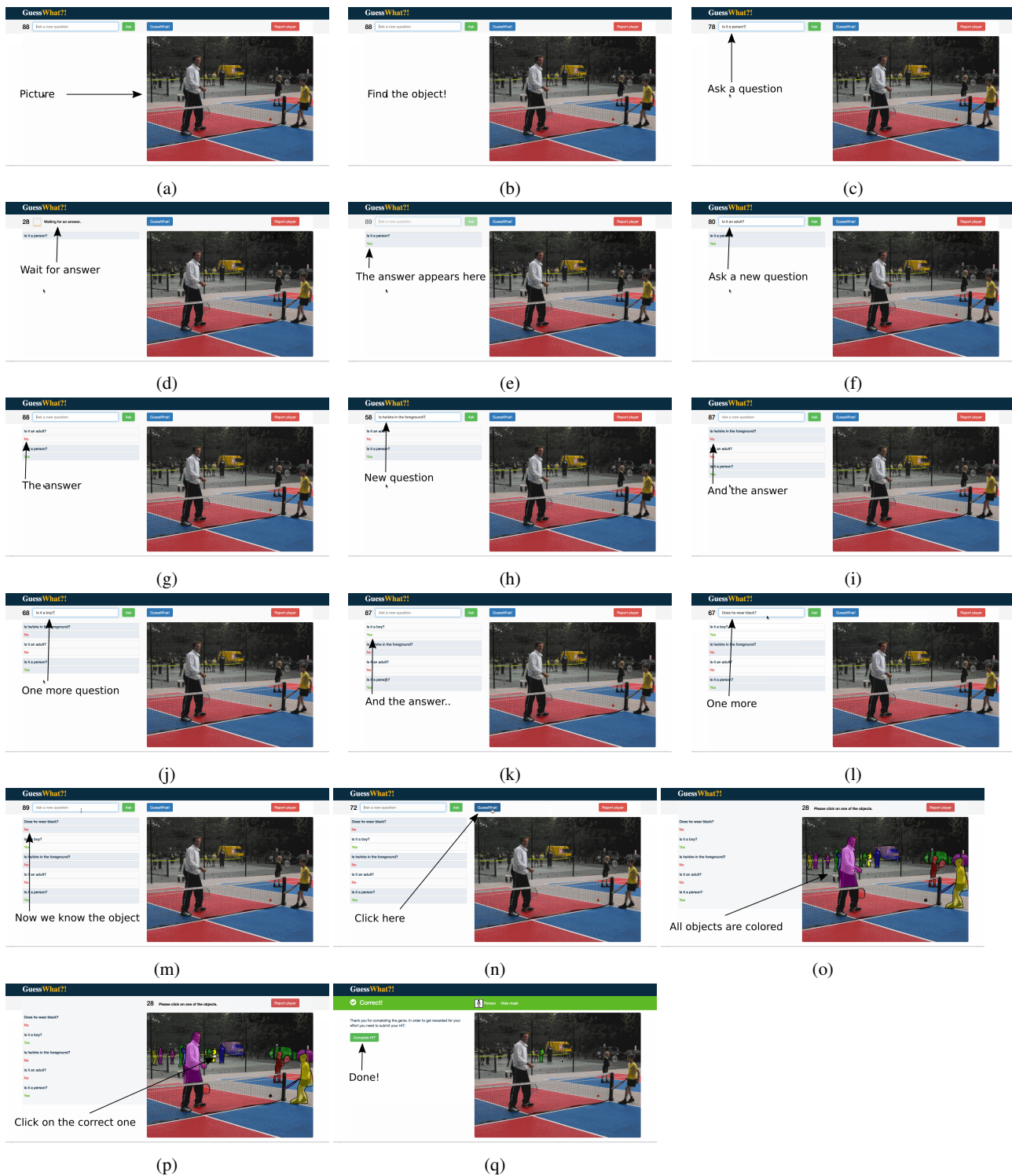


Figure 9: An example game from the perspective of the questioner. Shown from left to right and top to bottom.

GuessWhat?!

Instructions

You have to be a native english speaker in order to participate in this HIT.

Your task is to reformulate the following 25 questions which probably contain some spelling or grammar mistakes. The questions are extracted from the GuessWhat!? game in which a player locates an object in a picture by asking yes/no questions. For some questions more context is necessary to correct them, clicking on the 'show dialogue' button will show the game which the question was part of.

Two example corrections:

- ba na na? -> Is it a banana?
- in left of pic the far 2... one of them -> Is it one of the far two in the left of the picture?

Also make sure you retain the meaning of the question. Note further that a questions must:

- contain at least three words
- end with a question tag (?)

Your correction will be colored:

- In **red** if it does not follow the above constraints
- In **orange** if no correction has been made (in some cases this is fine)
- In **green** if it has been modified

Some displayed questions may have been corrected by other workers. If the displayed question makes no sense, you may display the original question by hovering it or displaying the full dialogue.

For any further questions regarding the HIT, please contact Harm de Vries at guesswhat.mturk@gmail.com.

Questions	Your correction	More context
a vehicle >		Display dialogue >
is it in firstbox		Display dialogue >
Do you see more than 3 clear pastic bottles on the top of the table?		Display dialogue >
the chapati .		Display dialogue >
2nd \ \		Display dialogue >
ia it on the left		Display dialogue >
ok I see something completely yellow (not orange). Is this thing next to the yellow		Display dialogue >

(a) Interface to fix ill-formatted questions

GuessWhat?!

Instructions

You have to be a native english speaker in order to participate in this HIT.

Your task is to check whether the following 50 questions were correctly reformulate. The questions are extracted from the GuessWhat!? game in which a player locates an object in a picture by asking yes/no questions. For some questions more context is necessary to correct them, clicking on the 'show dialogue' button will show the game which the question was part of.

You have to select

- **Yes** if the reformulation preserve the initial meaning of the question AND there is no english mistakes
- **No** when the reformulation lower the initial meaning of the question OR if the reformulation contains english mistakes
- **Report** if the original/final question make no-sense regarding the dialogue OR the final question contains slang words

Two example corrections:

- ba na na? -> Is it a banana? **Yes**
- in left of pic the far 2... one of them -> Is it one of the far two in the left of the picture? **Yes**
- Is it the man is the wite T-shirt? -> Is it a man? **No**
- Is it 1 of these f*** eggs? -> Is it one of these f***** eggs? **Report**
- Is it 1 of the f*** eggs? -> Is it one of the eggs? **Yes**

For any further questions regarding the HIT, please contact Harm de Vries at guesswhat.mturk@gmail.com.

Questions	Correction	Diff	Yes	No	Report	
Is it the car behind the bus (you can see all of it)	Is the car behind the bus?	Is it the car behind the bus (you can see all of it)?	☐	☐	☐	Display dialogue >
cp . . .	Is it the cap?	cp . . . Is it the cap?	☐	☐	☐	Display dialogue >
the whole (half) of a piece	Is it the whole or half of a piece?	Is it the whole (or half) of a piece?	☐	☐	☐	Display dialogue >
is it located at the left part of the pic>	Is it located at the left part of the picture?	Is it located at the left part of the pic> ture?	☐	☐	☐	Display dialogue >
is he wear white t-shirt?	Is he wearing a white t-shirt?	Is he wearing a white t-shirt?	☐	☐	☐	Display dialogue >
It is the person in white that we can only see their shudlers and head?	Is it the top half of a person wearing white?	It is it the person in white that we can only see their shudlers and head top half of a person wearing white?	☐	☐	☐	Display dialogue >
is it a doughnut?	Is it a doughnut?	Is it a doughnut?	☐	☐	☐	Display dialogue >
is it skatebard?	Is it a skateboard?	Is it a skatebaard?	☐	☐	☐	Display dialogue >

(b) Interface to validate the fix ill-formatted questions

Figure 10: In the first task, we ask workers to correct mistakes in the questions. We then ask workers to validate the proposed correction by showing the difference between the original question and its correction. We alternate both tasks till all questions are corrected and validated.

B. GuessWhat?! samples



Is it a person? **No**
 Is it a cat? **Yes**
 Is it in the hands of these girls? **No**
 The cat in the right side of the image? **Yes**



Is it human? **Yes**
 Are they five humans visible? **Yes**
 Is it leftmost? **No**
 Is it in the middle? **Yes**
 Is it the third one from the left? **Yes**



Is it an ear? **No**
 Is it a horn? **No**
 Is it a cow? **Yes**
 Is it the one at the back? **No**
 Does it have a horn? **Yes**

Figure 11: Three other examples of our dataset.



Is it a person? **Yes**
 Is it in the foreground? **No**
 Is he wearing blue shirt? **No**
 Is he wearing black shirt? **Yes**



Is it a person? **Yes**
 Is it standing? **Yes**
 Is it in yellow? **Yes**



Is it a person? **Yes**
 One in yellow? **No**
 One with white pants? **No**
 One with blue pants in the top? **No**

Figure 12: Same picture but different objects.



Is it a person? **No**
 Is it on the shelves? **No**
 Is it on the floor? **Yes**
 Is it a cup? **No**
 Is it blue? **No**
 Is it wood? **No**
 Is it the bed? **No**
 Is it in the lower half of the image? **Yes**
 Is it in the lower left corner? **No**
 Is it near the blue tray with two cups? **No**
 Is it near the green and blue toy? **No**
 Is it near the boy laying on the floor? **Yes**
 Is it something he is touching? **Yes**
 Is it the remote? **Yes**
 Is it the one in his right hand (close to the wooden box)? **No**

Figure 13: A long dialogue example in a very rich environment.



ReferIt

woman in red jacket with green bag
left woman in red coat

GuessWhat?!

Is it a person? **Yes**
One of the people with the stroller on the right? **No**
One of the two people crossing the street towards us? **No**
The woman in red? **Yes**

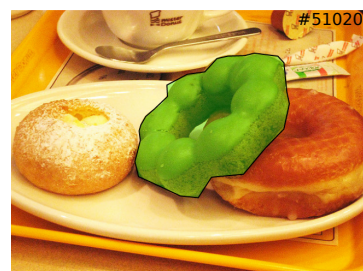


ReferIt

guy with hat bottom right front
guy sitting with hat bottom right

GuessWhat?!

Is it a person? **Yes**
Are they standing? **No**
Are they touching the frisbee? **No**
Are they holding a square thing? **Yes**
Black cap ? **Yes**



ReferIt

doughnut in the middle with green frosting

GuessWhat?!

Is it edible? **Yes**
Is it on oval plate? **Yes**
Is it green? **Yes**
The whole doughnut? **Yes**

Figure 14: Samples illustrating the difference between GuessWhat?! and ReferIt games. As both dataset are constructed on top of MS COCO, we picked identical objects (and images).

C. Additional database statistics

Figure 15 presents a word co-occurrence matrix of the GuessWhat?! dataset. Figure 16 and Figure 17a compares the object size and category distribution of GuessWhat?! with MS Coco.

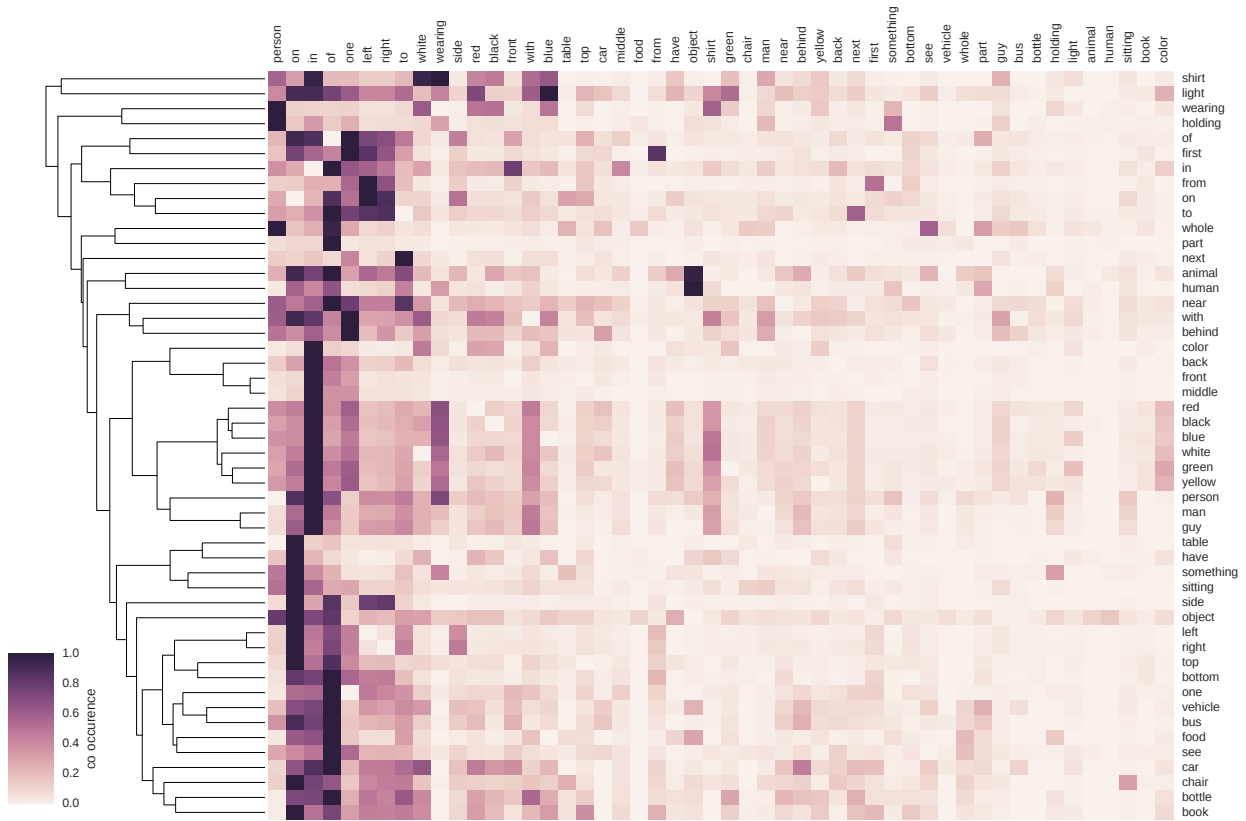


Figure 15: Co-occurrence matrix of words. Only the 50 most frequent words are kept. Rows are first normalized before being sorted thanks a hierarchical clustering with an euclidean distance.

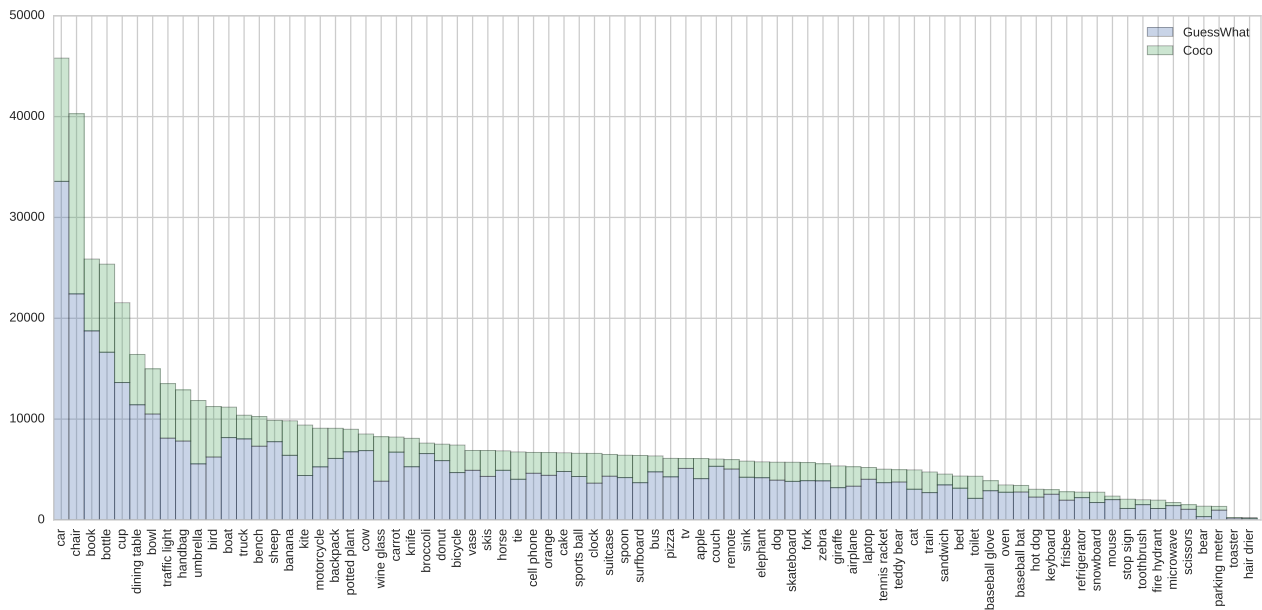
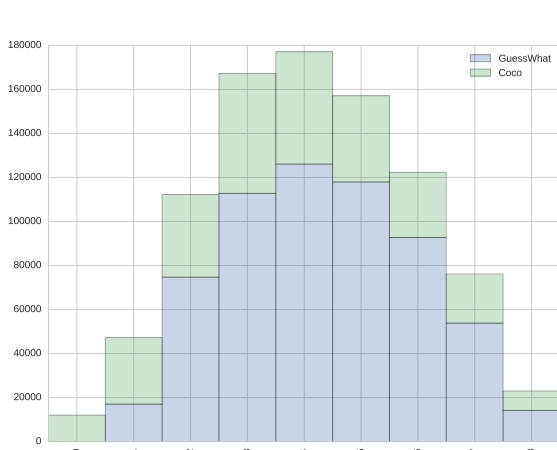
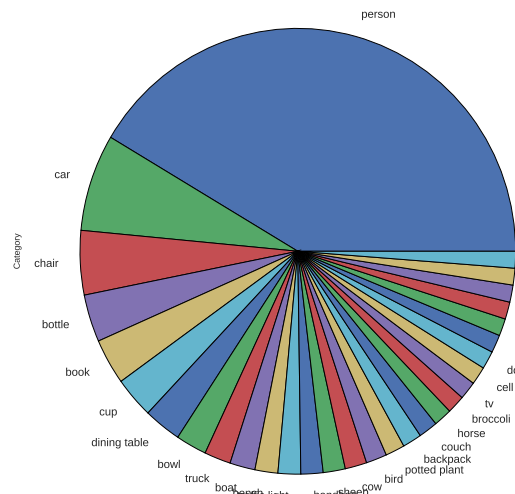


Figure 16: Visualization of the object category distribution of MS COCO and GuessWhat?! dataset. The person category was removed for clarity (resp. 273469 and 188204).



(a)



(b)

Figure 17: (a) Visualization of the object size distribution of MS COCO and GuessWhat?! dataset. (b) Distribution of the the 30 (out of 80) prominent object categories in the GuessWhat?! which represent 71.3% of the objects.

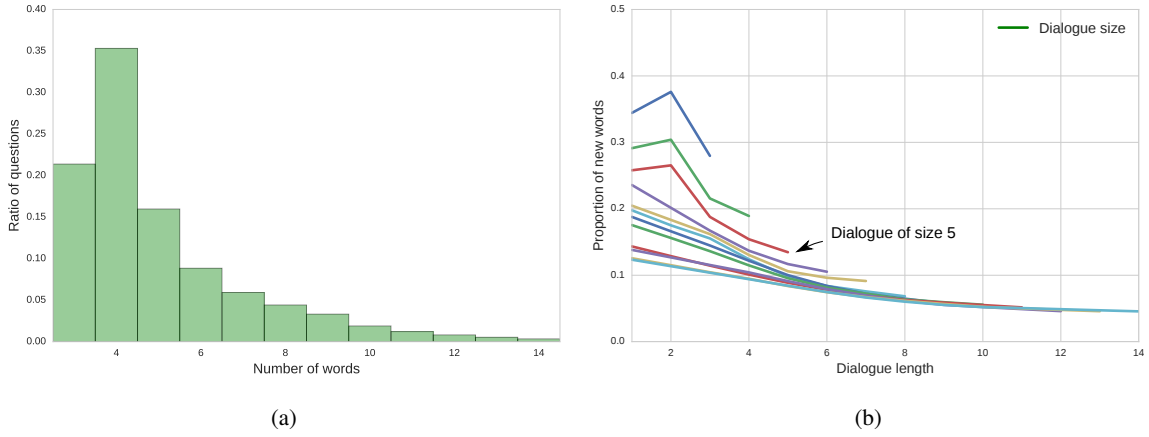


Figure 18: (a) Number of words per question. The question length follows a Poisson-like distribution, a finding which is in line with other datasets [6]. (b) Percentage of apparition of new words along a dialogues. Questioner tends keep using the same words during the dialogues.

Topic 1	Topic 2
Abstract words	Descriptive words
person	left
food	one
vehicle	right
human	wearing
car	side
one	white
object	red
animal	table

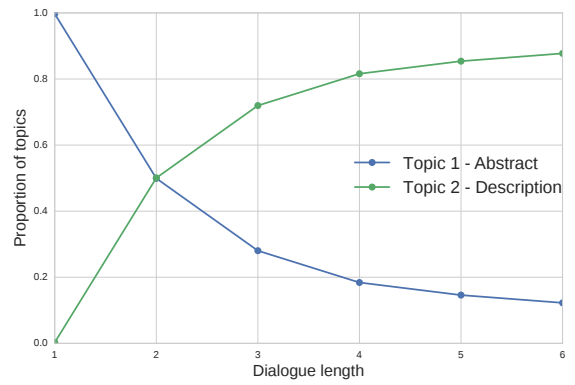


Figure 19: Relative evolution of topics during a dialogue of size 6. We applied Data Topic Models (DTM) [8] with the python framework [32] on our dataset. The table reports the two prominent detected topics with their respective key words while the figure display their relative evolution during the dialogue. The topic titles are manually picked.

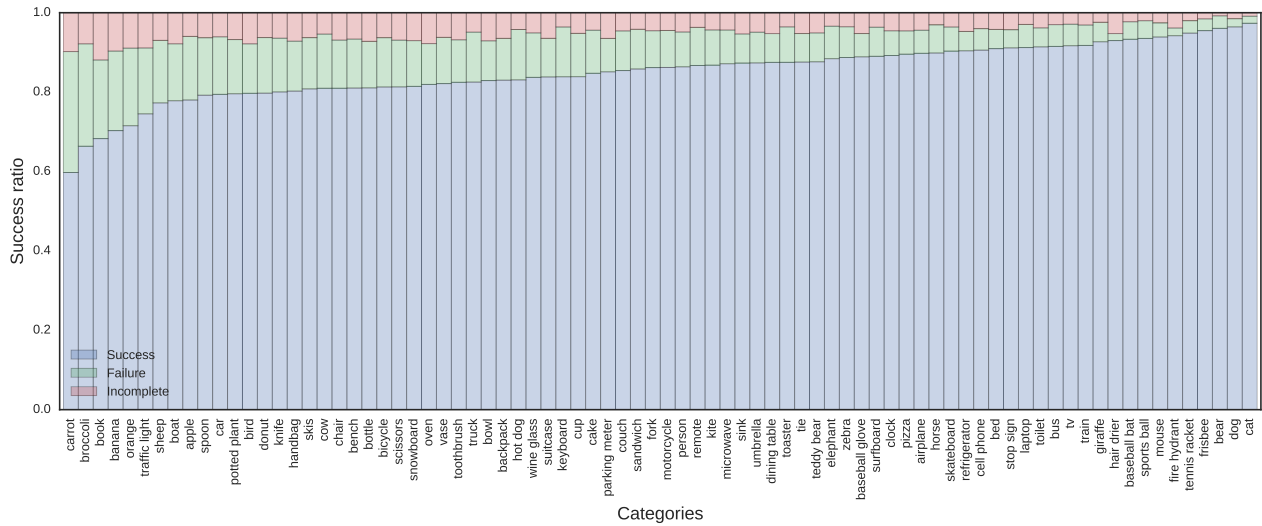


Figure 20: Histogram of success ratio broken down per object category.

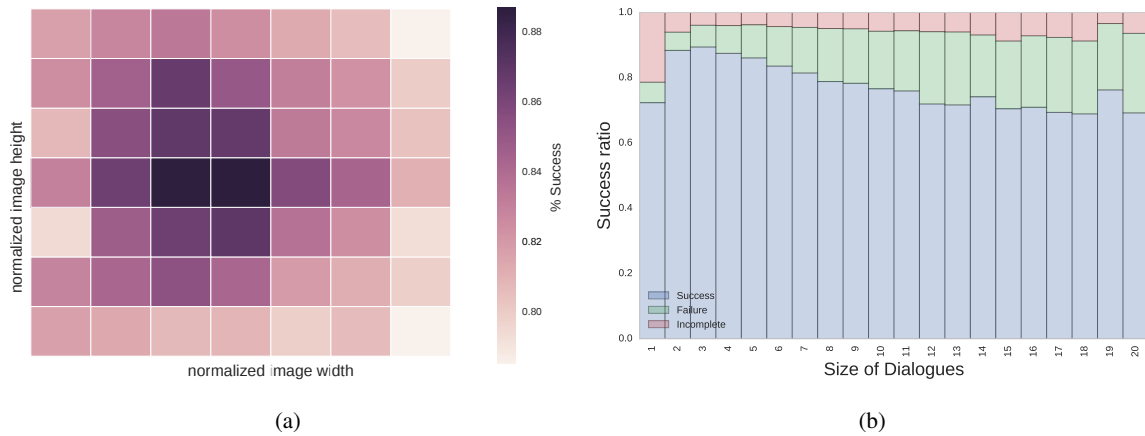


Figure 21: (a) Heatmap of the success ratio with respect to the spatial location within the picture. (b) Histogram of the success ratio relative to the dialogue length.

person	14.48	person	3.20	=	left	2.95	↗
food	1.29	left	1.69	↗	right	2.32	↗
animal	1.16	wearing	1.20	new	person	2.28	↗
human	1.03	right	1.02	new	wearing	1.66	↗
object	0.77	front	0.97	new	whole	1.58	new
car	0.60	white	0.91	new	white	1.56	=
vehicle	0.57	red	0.77	new	red	1.26	=
cat	0.41	car	0.64	↗	black	1.19	↗
alive	0.37	black	0.60	new	front	1.14	↗
dog	0.35	blue	0.59	new	blue	1.10	=

(a) Dialogues having 3 questions

person	8.20	person	1.98	=	left	1.92	=	left	2.13	=
food	1.03	left	1.03	↗	right	1.77	↗	right	2.04	=
human	0.56	right	0.66	new	person	1.20	↗	white	1.28	↗
animal	0.46	front	0.59	new	white	1.12	=	person	1.26	↗
vehicle	0.45	car	0.51	↗	wearing	0.93	=	black	0.90	↗
object	0.42	white	0.48	new	side	0.79	↗	wearing	0.85	↗
car	0.36	wearing	0.48	new	red	0.72	=	red	0.80	=
furniture	0.24	side	0.43	new	front	0.69	↗	whole	0.80	new
left	0.24	red	0.39	new	black	0.65	↗	blue	0.75	=
edible	0.20	vehicle	0.39	↗	blue	0.58	↗	front	0.73	=

(b) Dialogues having 5 questions

person	5.89	person	1.44	=	left	1.26	=	left	1.42	=	left	1.65	=
food	0.74	left	0.73	↗	right	1.08	↗	right	1.39	=	right	1.54	=
human	0.38	right	0.42	new	person	0.82	↗	white	0.88	=	white	0.96	=
vehicle	0.30	table	0.37	↗	side	0.67	↗	person	0.84	=	person	0.90	=
object	0.28	front	0.36	new	side	0.60	↗	black	0.63	↗	red	0.65	↗
car	0.26	food	0.35	↗	wearing	0.54	=	red	0.57	↗	black	0.63	↗
animal	0.26	side	0.35	new	red	0.49	=	wearing	0.56	↗	blue	0.57	↗
furniture	0.20	car	0.31	↗	table	0.41	=	blue	0.54	↗	wearing	0.52	↗
left	0.14	wearing	0.28	new	front	0.41	↗	side	0.53	↗	next	0.51	new
boat	0.14	something	0.28	new	car	0.37	↗	front	0.45	=	side	0.51	↗

(c) Dialogues having 7 questions

Table 5: Proportions of the ten most common words for each depth of questions for sorted by the size of the dialogues

D. All oracle baselines

Model	Train err	Valid err	Test err
Dominant class (no)	47.4%	46.2%	50.9%
Category	43.0%	42.8%	43.1%
Question	40.2%	41.7%	41.2%
Crop	40.9%	42.7%	43.0%
Image	45.7%	46.7%	46.7%
Spatial	43.9%	44.1%	44.3%
Category + Spatial	41.6%	41.7%	42.1%
Question + Crop	22.3%	29.1%	29.2%
Question + Image	37.9%	40.2%	39.8%
Question + Category	23.1%	25.8%	25.7%
Question + Spatial	28.0%	31.2%	31.3%
Spatial + Crop	41.8%	42.4%	42.8%
Crop + Image	41.6%	42.1%	42.4%
Spatial + Image	42.2%	44.1%	44.2%
Category + Crop	41.0%	41.7%	42.3%
Category + Image	42.3%	42.7%	43.0%
Category + Crop + Image	40.6%	41.5%	41.8%
Category + Spatial + Crop	40.6%	41.6%	42.1%
Question + Category + Spatial	17.2%	21.1%	21.5%
Question + Crop + Image	23.7%	29.9%	30.0%
Category + Spatial + Image	40.4%	42.0%	42.2%
Question + Category + Image	23.4%	27.1%	27.4%
Question + Spatial + Image	28.4%	32.5%	32.5%
Spatial + Crop + Image	41.6%	42.1%	42.5%
Question + Category + Crop	20.4%	24.4%	24.7%
Question + Spatial + Crop	19.4%	26.0%	26.2%
Question + Category + Spatial + Crop	16.1%	21.7%	22.1%
Question + Spatial + Crop + Image	20.7%	27.7%	27.9%
Category + Spatial + Crop + Image	40.3%	41.4%	41.8%
Question + Category + Spatial + Image	19.2%	23.2%	23.5%
Question + Category + Crop + Image	20.0%	25.3%	25.5%
Question + Category + Spatial + Crop + Image	17.8%	23.2%	23.3%

Table 6: Classification errors for all oracle baselines.

E. Guesser generation model

	Generated		Groundtruth	
	Is it a person?	No	Is it in the sky?	No
	Is it the kite?	No	Is it the umbrella?	No
	Is it the kite?	No	Is it the ocean?	No
	Is it the chair?	No	Is it the boat?	Yes
	Is it the boat?	Yes		
	Is it a person?	No	Is it an object?	Yes
	Is it a skateboard?	No	Do you wear it?	No
	Is it a car?	Yes	Do you ride it?	No
	Is it the one on the right?	No	Is it metal?	Yes
	Is it the one on the right?	No		
	Is it a person?	Yes	Is it a person?	Yes
	Is it the one in the front?	No	Is he in the foreground?	No
	Is it the one in the middle?	Yes	Is he wearing blue?	Yes
	Is it the one in the middle?	Yes		
	Is it the whole person?	Yes		

Figure 22: Three samples of QGen+GT model for which the correct object was predicted.

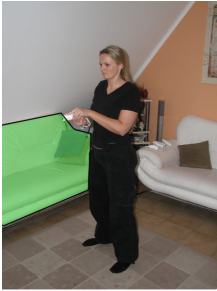
	Generated	Groundtruth
	Is it the cat? No Is it the cat? No Is it the chair? No Is it the book? No Is it the book? No	Is it an animal? No Is it a device? Yes Is it silver in color? Yes
	Is it a person? No Is it a remote? No Is it the chair? Yes Is it the one on the right? No Is it the one next to the right? No	Is it a person? No Is it a couch? Yes Does the couch have two pillows on it? Yes
	Is it a person? Yes Is it the guy in the front? No Is it the guy in the middle? No Is it the guy in the middle? No Is it the guy in the middle? No	Is it a person? Yes Is it in the foreground? No Is it on a screen? Yes

Figure 23: Three dialogue samples of QGen+GT model for which the wrong object was predicted.