#### A. User interface

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

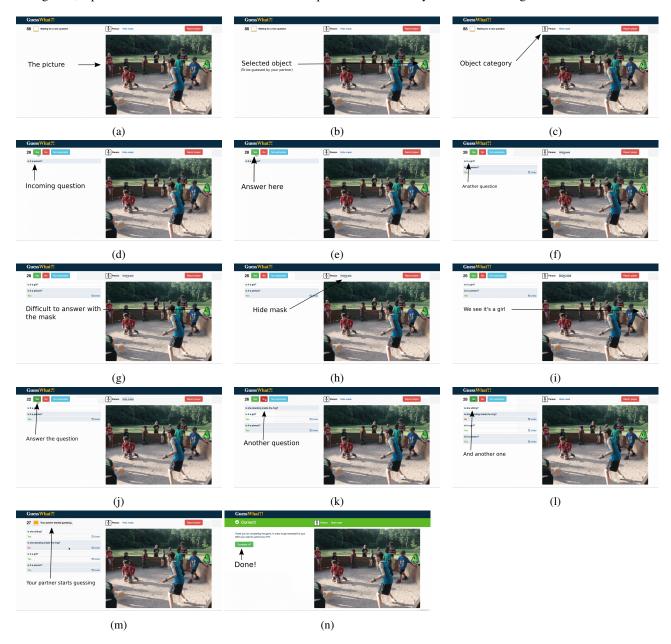


Figure 8: An example game from the perspective of the oracle. Shown from left to right and top to bottom.

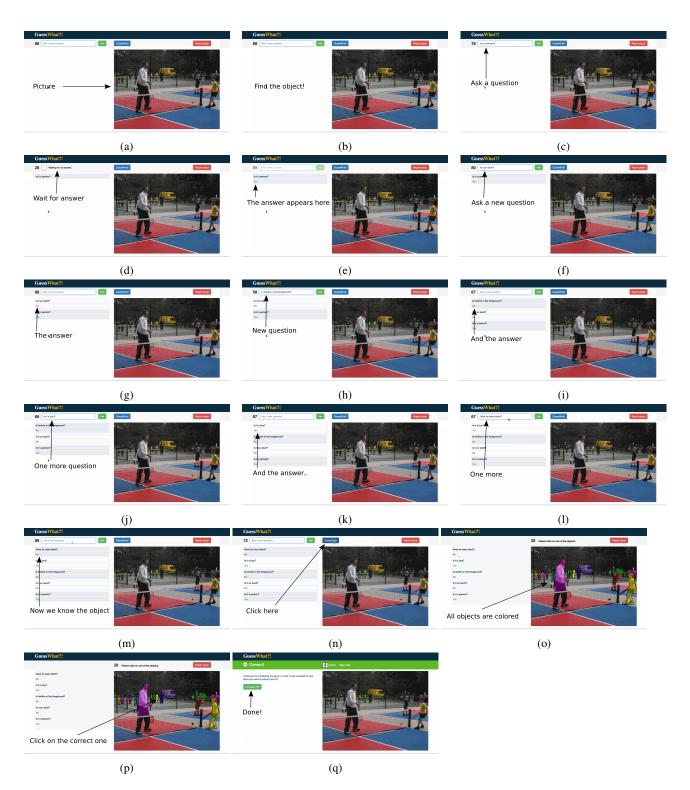


Figure 9: An example game from the perspective of the questioner. Shown from left to right and top to bottom.

nstructions						
u have to be a native english speaker in ord	er to participate in this HIT.					
ur task is to reformulate the following 25 que nich a player locates an object in a picture by ow the game which the question was part of	asking yes/no questions. For some					
vo example corrections:						
<ul> <li>ba na na? -&gt; Is it a banana?</li> <li>in left of pic the far 2 one of them -&gt; Is</li> </ul>	it one of the far two in the left of the	e picture?				
on make sure you retain the meaning of the contain at least three words	question. Note further that a question	ons must:				
end with a question tag (?)  our correction will be colored:						
In red if it does not follow the above cons     In orange if no correction has been mad     In green if it has been modified						
ome displayed questions may have been cor	rected by other workers. If the displ	ayed question makes no sens	se, you may display the o	riginal q	uestion by h	overing it or displaying
e full dialogue. or any further questions regarding the HIT, pl	ease contact Harm de Vries at gues	sswhat.mturk@gmail.com.				
Questions			our correction		More	e context
vehicle >					Dis	splay dialogue »
s it in firstbox					Dis	splay dialogue »
Do you see more than 3 clear pastic bottles o	n the top of the table?					splay dialogue »
he chapati .					Dis	splay dialogue »
2nd \ \					Dis	splay dialogue »
ia it on the left					Die	splay dialogue »
k I see something completely yellow (not ora	inge). Is this thing next to the yellow	V			Dis	splay dialogue »
GuessWhat?!	(a) Interface to f	nx ill-formatted	questions			
netructions						
	dar to participate in this LHT					
fou have to be a native english speaker in ord four task is to check weither the following 50 picture by asking yes/no questions. For some	questions were correctly reformulat					
You have to be a native english speaker in ord Your task is to check weither the following 50 picture by asking yes/no questions. For some part of.	questions were correctly reformulat					
fou have to be a native english speaker in ord four task is to check weither the following 50- sicture by asking yes/no questions. For some art of.  • You have to select • Yes if the reformulation preserve the ini • No when the reformulation lower the ini • Report if the original/final question mak	questions were correctly reformulat questions more context is necessa tial meaning of the question AND th tial meaning of the question OR if the	ry to correct them, clicking on nere is no english mistakes he reformulation contains eng	the 'show dialogue' butt			
You have to be a native english speaker in or four task is to check weither the following 50 - licture by asking yes/no questions. For some art of.  You have to select  Yes if the reformulation preserve the in in No when the reformulation lower the in Report if the original/final question mak (we example corrections:	questions were correctly reformulat questions more context is necessa tial meaning of the question AND th tial meaning of the question OR if the	ry to correct them, clicking on nere is no english mistakes he reformulation contains eng	the 'show dialogue' butt			
fou have to be a native english speaker in or four task is to check weither the following 50 circlure by asking yes/no questions. For some and of.  Ves if the reformulation preserve the ini No when the reformulation lower the ini Report if the original/final question mak five example corrections.  • ba na na? -> Is it a banana? Yes • in left of pic the far Z one of them -> Is	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if the e no-sense regarding the dialogue is it one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the far two in the left of the sit one of the sit one sit one sit one of the sit one sit one sit one one	ny to correct them, clicking on the clicking on the reformulation contains eng OR the final question contains	the 'show dialogue' butt			
ou have to be a native english speaker in or four task is to check weither the following 50 icture by asking yes/no questions. For some art of.  four have to select  Yes if the reformulation preserve the ini No when the reformulation lower the ini Report if the original/linal question mak wo example corrections:  ba na na? -> Is it a banana? Yes  in left of pic the far 2 one of them -> Is is it the man is the wite T-shirt? -> Is it	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if the eno-sense regarding the dialogue is it one of the far two in the left of the a man? No	ny to correct them, clicking on the clicking on the reformulation contains eng OR the final question contains	the 'show dialogue' butt			
ou have to be a native english speaker in or four task is to check weither the following 50 iciture by asking yes/no questions. For some art of.  You have to select  Yes if the reformulation preserve the in in Nowhen the reformulation lower the ining the original final question mak wise example corrections:  bana na? > Is it a banana? Yes in left of pic the far Z one of them > It is it the man is the wite T-shirt? >> Is it is it of the eft' eggs? -> Is it one of the of its it of the fire eggs? -> Is it no of the	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if the no-sense regarding the dialogue is it one of the far two in the left of the a man? No these from eggs? Report e eggs? Yes	ry to correct them, clicking on here is no english mistakes he reformulation contains eng OR the final question contains the picture? Yes	the 'show dialogue' butt			
ou have to be a native english speaker in or our task is to check weither the following 50 cicture by asking yes/no questions. For some art of.  ou have to select  • Yes if the reformulation preserve the in in  • No when the reformulation lower the ini  • Report if the original/final question mak wo example corrections:  • ba na na? -> is it a banana? Yes  • in left of pic the far 2 one of them -> is  • is it the man is the wite T-shirt? -> js it is  • is it 1 of the set ""eggs? -> ) is it one of the of its it 1 of the it" eggs? -> ) is it no of the of	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if the no-sense regarding the dialogue is it one of the far two in the left of the a man? No these from eggs? Report e eggs? Yes	ry to correct them, clicking on here is no english mistakes he reformulation contains eng OR the final question contains the picture? Yes	the 'show dialogue' butt			
ou have to be a native english speaker in ord our task is to check weither the following 50 clure by asking yes/no questions. For some art of.  ou have to select  • Yes if the reformulation preserve the inity of the reformulation lower the inity experience. The reformulation lower the inity experience in the reformulation lower the inity experience in the reformulation make the selection of the reformulation was example corrections:  • ba na na? -> Is it a banana? Yes  • in left of pic the far 2 one of them -> Is  • is it the man is the wift = Tshift? -> Is it  • is it if an is the wift = Tshift? -> Is it one of the ord of the	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if the no-sense regarding the dialogue is it one of the far two in the left of the a man? No these from eggs? Report e eggs? Yes	ry to correct them, clicking on here is no english mistakes he reformulation contains eng OR the final question contains the picture? Yes	the 'show dialogue' butt		now the gam	
ou have to be a native english speaker in or our task is to check weither the following 50 citure by asking yes/no questions. For some art of.  bu have to select  • Yes if the reformulation preserve the in in No when the reformulation lower the in Pepor if the original/final question mak wo example corrections:  • ba na na? > Is it a banana? Yes • in left of pic the far 2 one of them > Ie is it 1 of these f" eggs? -> Is it one of is it 1 of these f" eggs? -> Is it one of the or any further questions regarding the HIT, p	questions were correctly reformulat questions more context is necessal tital meaning of the question AND that meaning of the question OR if the necessary of the question or in the left of the necessary of the dialogue of the far two in the left of the near near near near near near near nea	ury to correct them, clicking on the control of the	the 'show dialogue' butt	on will sh	o Report	
ou have to be a native english speaker in or our task is to check weither the following 50 cture by asking yes/no questions. For some art of.  • Yes if the reformulation preserve the ini • No when the reformulation lower the ini • Report if the original/final question mak wo example corrections: • ba na na? -> Is it a banana? Yes • in let of pic the far 2 one of them -> Is • is if the man is the wite T-shirt? -> Is it is • is if 1 of these ("reggs? -> ) is it one of the or any further questions regarding the HIT, p	questions were correctly reformulat questions more context is necessal tital meaning of the question AND the tital meaning of the question OR if it encountered to the non-sense regarding the dialogue is it one of the far two in the left of the a man? No these from eggs? Report e eggs? Yes lelease contact Harm de Vries at que Correction	ury to correct them, clicking on the control of the	the 'show dialogue' butt	Yes N	o Report	e which the question v
ou have to be a native english speaker in ord our task is to check weither the following 50 cicture by asking yes/no questions. For some art of.  • Yes if the reformulation preserve the initial value of the ordinal valu	questions were correctly reformulat questions more context is necessal tital meaning of the question AND that tital meaning of the question OR if it is no eof the far two in the left of the a man? No these trimeges? Report e eggs? Yes lease contact Harm de Vries at que Correction  Is the car behind the bus?	ury to correct them, clicking on the control of the	the 'show dialogue' butt	Yes N	o Report	e which the question v
ou have to be a native english speaker in or our task is to check weither the following 50 cicture by asking yes/no questions. For some art of.  • Yes if the reformulation preserve the init on the way of the worker than the properties of the worker than the properties of the properties of the worker than the worker t	questions were correctly reformulat questions more context is necessal tital meaning of the question AND the lital meaning of the question OR if it is necessary to the question OR if it is necessary to the question OR if it is not of the far two in the left of the aman? No these from eggs? Report eggs? Yes lease contact Harm de Vries at guestion of the properties of the car behind the bus?	ury to correct them, clicking on the control of the	the 'show dialogue' butt  lish mistakes s slang words  (you can see all of it)?	Yes N	o Report	Display dialogue -
ou have to be a native english speaker in ord our task is to check weither the following 50 cicture by asking yes/no questions. For some art of.  • Yes if the reformulation preserve the initial value of the original/line or them or in the original/line	questions were correctly reformulat questions more context is necessal tital meaning of the question AND the tital meaning of the question OR if if the nearing of the question OR if if the near of the far two in the left of the near of the far two in the left of the near of the far two in the left of the near of the	ury to correct them, clicking on the control of the	the 'show dialogue' but  lish mistakes s slang words  (you can see all of it)?	Yes N	o Report	Display dialogue - Display dialogue - Display dialogue -
ou have to be a native english speaker in ord our task is to check weither the following 50 cicture by asking yes/no questions. For some art of.  • Yes if the reformulation preserve the init ou have to select  • Yes if the reformulation preserve the init on the order that in the original/linal question make we example corrections:  • ba nan ana? > Is it a banana? Yes  • in let of pic the far 2 no end them > Is  • is it the man is the wite T-shirt? -> Is it or in the original/linal questions are in let of pic the far 2 no end them > Is  • is at 1 of thee? "" eggs? -> Is it one of the or any further questions regarding the HIT, proceeding the HIT, procee	questions were correctly reformulat questions more context is necessal tital meaning of the question AND that meaning of the question OR if it is no end the far two in the left of that man? No these f** ggs? Report e eggs? Yes elease contact Harm de Vries at que the car behind the bus?  Let the car behind the bus?  Is it the cap?  Is it the whole or half of a piece?  Is it located at the left part of the picture?	ury to correct them, clicking on the control of the	the 'show dialogue' but  lish mistakes s slang words  (you can see all of it)?  piece? of the pic>ture?	Yes N	o Report	Display dialogue - Display dialogue - Display dialogue - Display dialogue -
No when the reformulation lower the inity Report if the original/final question make Two example corrections:  I have a na? > Is it a banana? Yes  in left of pic the far 2 one of them -> Is  it the man is the wife T-shirt? -> Is it one of the series	questions were correctly reformulat questions more context is necessal tial meaning of the question AND the tial meaning of the question OR if it enosense regarding the dialogue is it one of the far two in the left of the aman? No these the caps? Seport e eggs? Report e eggs? Report e eggs? Report lease contact Harm de Vries at gue Correction  Is the car behind the bus?  Is it the whole or half of a piece?  Is it located at the left part of the picture?  Is he wearing a white t-shirt?	ury to correct them, clicking on the control of the	the 'show dialogue' but  lish mistakes s slang words  (you can see all of it)?  piece? of the pic>ture?	Yes N	o Report	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



Yes No Yes

Is it a person?
Is it a cat?
Is it in the hands of these girls?
The cat in the right side of the
image?



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.



### <u>ReferIt</u>

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?
Are they standing?
Are they touching the frisbee
Are they holding a square
thing?
Black cap ?



#### <u>ReferIt</u>

doughnut in the middle with green frosting

#### **GuessWhat?!**

Yes

Yes	Is it edible?	Yes
No	Is it on oval plate?	Yes
No	Is it green?	Yes
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.

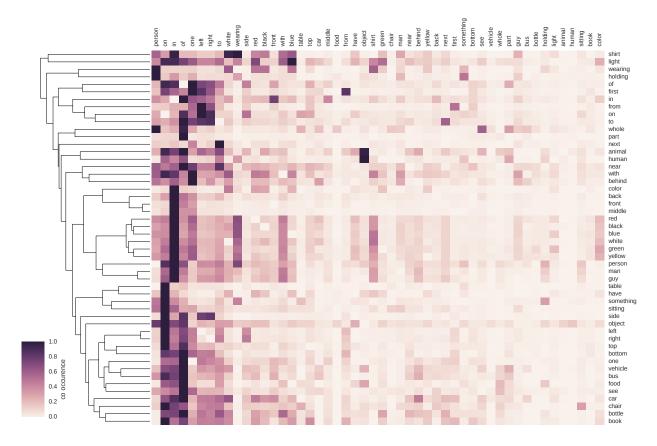


Figure 15: Co-occurrence matrix of words. Only the 50 most frequent words are kept. Rows are first normalized before being sorted thanks a hierarchical clustering with an euclidean distance.

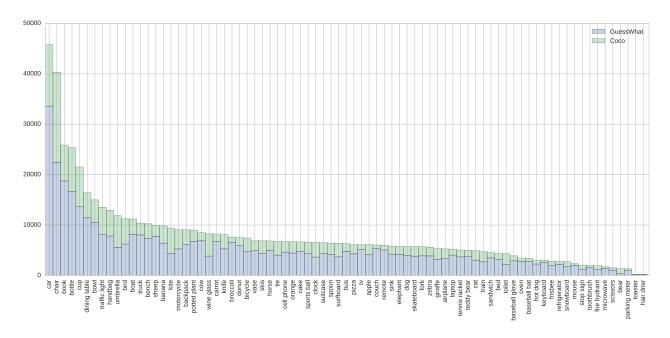


Figure 16: Visualization of the object category distribution of MS COCO and GuessWhat?! dataset. The person category was removed for clarity (resp. 273469 and 188204).

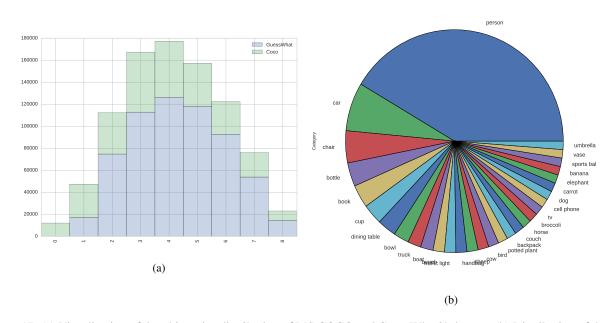


Figure 17: (a) Visualization of the object size distribution of MS COCO and GuessWhat?! dataset. (b) Distribution of the the 30 (out of 80) prominent object categories in the GuessWhat?! which represent 71.3% of the objects.

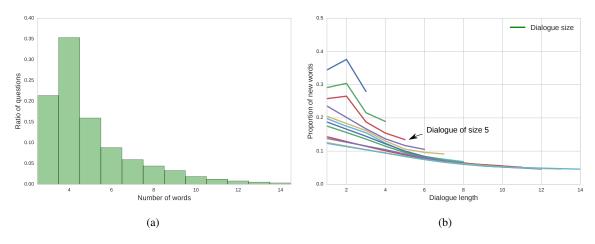


Figure 18: (a) Number of words per question. The question length follows a Poisson-like distribution, a finding which is in line with other datasets [6]. (b) Percentage of apparition of new words along a dialogues. Questioner tends keep using the same words during the dialogues.

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

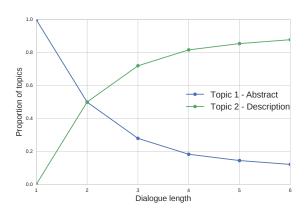


Figure 19: Relative evolution of topics during a dialogue of size 6. We applied Data Topic Models (DTM) [8] with the python framework [32] on our dataset. The table reports the two prominent detected topics with their respective key words while the figure display their relative evolution during the dialogue.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

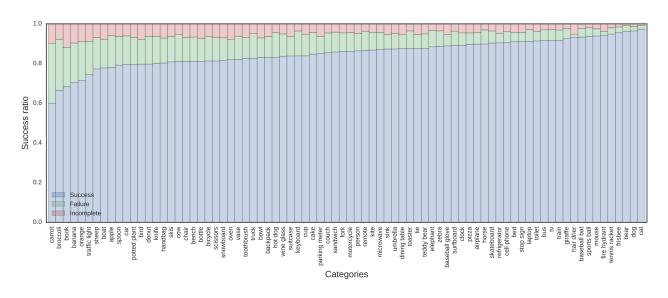


Figure 20: Histogram of success ratio broken down per object category.

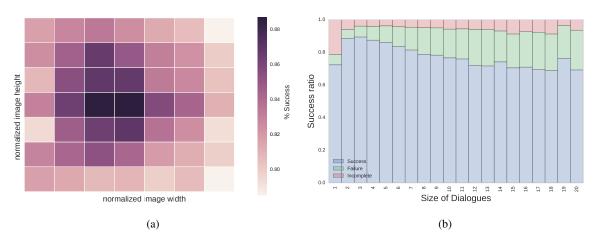


Figure 21: (a) Heatmap of the success ratio with respect to the spatial location within the picture. (b) Histogram of the success ratio relative to the dialogue length.

II	1.10	onne	IICM	0.39	anio	0.33	gon
٦	-	1.1			1.1	400	-
<u></u>	1.14	front	new	09.0	black	0.37	alive
	1.19	black	7	0.64	car	0.41	cat
II	1.26	red	new	0.77	red	0.57	vehicle
II	1.56	white	new	0.91	white	09.0	car
new	1.58	whole	new	0.97	front	0.77	object
ブ	1.66	wearing	new	1.02	right	1.03	human
7	2.28	person	new	1.20	wearing	1.16	animal
_	2.32	right		1.69	left	1.29	pooj
_	2.95	left	II	3.20	person	14.48	person

questions
$\alpha$
having
Dialogues
(a)

	_									
	ш	П	_	7	_	7	II	new	II	II
	2.13	2.04	1.28	1.26	0.60	0.85	0.80	0.80	0.75	0.73
- I	left	right	white	person	black	wearing	red	whole	blue	front
0	Ш	Κ_	7	П	II	Κ_	II	7	_	7
	1.92	1.77	1.20	1.12	0.93	0.79	0.72	69.0	0.65	0.58
	left	right	person	white	wearing	black	red	side	blue	front
	K	7	II	Κ_	_	Κ_	_	7	new	new
	1.69	1.41	1.26	0.84	0.82	0.67	0.62	0.58	0.55	0.54
	left	person	right	white	wearing	side	red	front	black	blue
	II	٢	new	new	٢	new	new	new	new	7
	1.98	1.03	99.0	0.59	0.51	0.48	0.48	0.43	0.39	0.39
	person	left	right	front	car	white	wearing	side	red	vehicle
	8.20	1.03	0.56	0.46	0.45	0.42	0.36	0.24	0.24	0.20
	person	food	human	animal	vehicle	object	car	furniture	left	edible

(b) Dialogues having 5 questions

person         5.89         person         1.44         =         left         1.26         =         left         1.25         =         left         1.39         =         left         1.42         =         left         1.26         =         left         1.26         =         left         1.26         =         left         1.39         =         left         1.54           human         0.38         right         0.42         new         right         0.89         =         person         0.82         xwiite         0.81         xwiite         0.88         =         right         1.54           vehicle         0.30         table         0.57         xwiite         0.67         xwiite         0.81         xwiite         0.88         =         right         1.54         =         right         1.59         =										_	
5.89         person         1.44         =         left         1.08         left         1.26         =         left         1.33         =         left         1.42         =           0.74         left         0.73         2         person         0.96         1eft         1.26         1eft         1.33         =         left         1.39         =           0.38         right         0.73         2         person         0.96         2         right         1.22         1right         1.39         =           0.30         table         0.57         3         white         0.60         3         white         0.88         =           0.28         front         0.35         washine         0.50         3         white         0.60         3         white         0.84         =           0.26         frond         0.35         washine         0.50         3         wearing         0.54         =         red         0.57         3           0.20         side         0.35         washine         0.36         4         red         0.49         =         red         0.54         3           0.20		1.65	1.54	96.0	0.90	0.65	0.63	0.57	0.52	0.51	0.51
5.89         person         1.44         =         left         1.08         left         1.26         left         1.33         =         left           0.74         left         0.73         person         0.96         left         1.26         left         1.33         left           0.38         right         0.73         person         0.96         left         1.08         left         1.22         left           0.30         table         0.37         left         0.89         left         1.08         left         1.22         left           0.28         table         0.37         left         0.89         left         0.80         left         left         left         left         0.81         left         left <td></td> <td>left</td> <td>right</td> <td>white</td> <td>person</td> <td>red</td> <td>black</td> <td>blue</td> <td>wearing</td> <td>next</td> <td>side</td>		left	right	white	person	red	black	blue	wearing	next	side
5.89         person         1.44         =         left         1.08         left         1.26         left         1.33         =         left           0.74         left         0.73         person         0.96         left         1.26         left         1.33         left           0.38         right         0.73         person         0.96         left         1.08         left         1.22         left           0.30         table         0.37         left         0.89         left         1.08         left         1.22         left           0.28         table         0.37         left         0.89         left         0.80         left         left         left         left         0.81         left         left <td></td> <td>II</td> <td>II</td> <td>II</td> <td>II</td> <td>Κ_</td> <td>Κ,</td> <td>7</td> <td>Κ_</td> <td>7</td> <td>II</td>		II	II	II	II	Κ_	Κ,	7	Κ_	7	II
5.89         person         1.44         =         left         1.08         left         1.26         =         left         1.33         =           0.74         left         0.73         left         1.08         left         1.26         left         1.33         =           0.38         right         0.77         left         1.08         left         1.25         =           0.30         table         0.37         left         1.08         left         1.25         =           0.28         right         0.89         left         1.08         left         1.22         left           0.29         table         0.57         left         1.08         left         1.25         left           0.26         table         0.57         left         left         left         0.81         left           0.26         frond         0.35         left         left <td></td> <td>1.42</td> <td>1.39</td> <td>0.88</td> <td>0.84</td> <td>0.63</td> <td>0.57</td> <td>0.56</td> <td>0.54</td> <td>0.53</td> <td>0.45</td>		1.42	1.39	0.88	0.84	0.63	0.57	0.56	0.54	0.53	0.45
5.89         person         1.44         =         left         1.08         1.26         =         left           0.74         left         0.73         2         person         0.96         2         right         1.08         2         right           0.38         right         0.42         new         right         0.82         2         white         right           0.28         front         0.37         3         side         0.57         3         white         0.67         3         white           0.26         front         0.35         we         wearing         0.48         3         wearing         0.54         =         side           0.26         side         0.35         w         red         0.41         new         red         0.49         =         red           0.20         car         0.31         1         table         0.39         1         table         0.41         =         red           0.14         wearing         0.28         w         front         0.38         1         black         0.41         3         blue           0.14         something         0.28		left	right	white	person	black	red	wearing	blue	side	front
5.89         person         1.44         =         left         1.08         1.26         =         left           0.74         left         0.73         2         person         0.96         2         right         1.08         2         right           0.38         right         0.42         new         right         0.82         2         white         right           0.28         front         0.37         3         side         0.57         3         white         0.67         3         white           0.26         front         0.35         we         wearing         0.48         3         wearing         0.54         =         side           0.26         side         0.35         w         red         0.41         new         red         0.49         =         red           0.20         car         0.31         1         table         0.39         1         table         0.41         =         red           0.14         wearing         0.28         w         front         0.38         1         black         0.41         3         blue           0.14         something         0.28		II	II	Κ_	7	Κ_	7	II	Κ,	Κ,	Κ_
5.89         person         1.44         =         left         1.08         left         1.26		1.33	1.22	0.81	0.80	0.59	0.57	0.54	0.51	0.49	0.42
5.89         person         1.44         =         left         1.08         left         1.26           0.74         left         0.73         left         1.08         left         1.26           0.38         right         0.96         left         1.26           0.30         table         0.37         left         1.08           0.28         front         0.36         wearing         0.82         white         0.67           0.26         front         0.35         wearing         0.48         left         0.60           0.20         side         0.35         left         0.49         left         0.49           0.20         car         0.31         left         1.48         left         0.49           0.20         car         0.31         left         0.39         left         1.49           0.14         wearing         0.28         new         front         0.39         left         0.41           0.14         wearing         0.28         new         front         0.37         left         0.41		left	right	white	person	wearing	side	red	black	blue	front
5.89         person         1.44         =         left         1.08         left           0.74         left         0.73         left         1.08         left           0.38         right         0.96         left         right           0.30         table         0.37         left         left           0.28         front         0.36         left         left           0.26         front         0.35         left         left           0.26         side         0.56         left         left           0.20         side         0.35         left         left           0.20         car         0.31         left         left           0.20         car         0.31         left         left           0.14         wearing         0.28         left         left           0.14         something         0.28         left         left           0.14         something         0.28         left         left		II	_	7	_	7	П	II	II	٢	new
5.89 person 1.44 = left 1.08	0	1.26	1.08	0.82	0.67	0.60	0.54	0.49	0.41	0.41	0.37
5.89         person         1.44         =         left         1.08           0.74         left         0.73         ↑         person         0.96           0.38         right         0.42         new         right         0.89           0.28         front         0.36         new         white         0.57           0.26         food         0.35         new         red         0.41         1           0.20         car         0.31         ↓         table         0.39         0.41         1           0.14         wearing         0.28         new         front         0.38         0.37           0.14         something         0.28         new         car         0.37		left	right	person	white	side	wearing	red	table	black	blue
5.89         person         1.44         =         left           0.74         left         0.73		K	7	II	Κ_	_	_	new	7	7	7
5.89 person 1.44 = 0.74 left 0.73		1.08	96.0	0.89	0.57	0.50	0.48	0.41	0.39	0.38	0.37
5.89         person         1.44           0.74         left         0.73           0.38         right         0.42           0.30         table         0.37           0.28         front         0.36           0.26         side         0.35           0.20         car         0.31           0.14         wearing         0.28           0.14         something         0.28		left	person	right	side	white	wearing	red	table	front	car
5.89 person 0.74 left 0.38 right 0.30 table 0.28 front 0.26 food 0.26 side 0.20 car 0.14 wearing 0.14 something		Ш	_	new	_	new	7	new	7	new	new
		1.4	0.73	0.42	0.37	0.36	0.35	0.35	0.31	0.28	0.28
		person	left	right	table	front	pooj	side	car	wearing	something
		5.89	0.74	0.38	0.30	0.28	0.26	0.26	0.20	0.14	0.14

(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? Is it the kite? Is it the kite? Is it the chair? Is it the boat?	No No No No Yes	Is it in the sky? Is it the umbrella? Is it the ocean? Is it the boat?	No No No Yes
WE THE THE THE THE THE THE THE THE THE TH				
	Is it a person? Is it a skateboard? Is it a car? Is it the one on the right? Is it the one on the right?	No No Yes No No	Is it an object? Do you wear it? Do you ride it? Is it metal?	Yes No No Yes
	Is it a person? Is it the one in the front? Is it the one in the middle! Is it the one in the middle! Is it the whole person?		Is it a person? Is he in the foregrou Is he wearing blue?	Yes nd? No 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 chair? No Is it the book? No Is it the book? No		Is it an animal? No Is it a device? Ye Is it silver in color? Ye	s
Is it a person? Is it a remote? Is it the chair? Is it the one on the right? Is it the one next to the right?	No No Yes No No	Is it a person? Is it a couch? Does the couch have two pillows on it?	No Yes Yes
Is it a person? Is it the guy in the front? Is it the guy in the middle? Is it the guy in the middle? Is it the guy in the middle?	Yes No No No No	Is it a person? Is it in the foreground? Is it on a screen?	Yes No Yes

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