

Region under Discussion for visual dialog

Mauricio Mazuecos^{1,2} and Franco Luque^{1,2} and Jorge Sánchez²
Hernán Maina^{1,2} and Thomas Vadora¹ and Luciana Benotti^{1,2}

¹Universidad Nacional de Córdoba

²CONICET, Argentina

{mmazuecos, hernan.maina, thvadora}@mi.unc.edu.ar
{francolq, jorge.sanchez, luciana.benotti}@unc.edu.ar

Abstract

Visual Dialog is assumed to require the dialog history to generate correct responses during a dialog. However, it is not clear from previous work how dialog history is needed for visual dialog. In this paper we define what it means for visual questions to require dialog history and we propose a methodology for identifying them. We release a subset of the Guesswhat?! questions for which their dialog history completely changes their responses. We propose a novel *interpretable* representation that visually grounds dialog history: the *Region under Discussion*. It constrains the image’s spatial features according to a semantic representation of the history inspired by the information structure notion of *Question under Discussion*. We evaluate the architecture on task-specific multimodal models and the visual transformer model LXMERT and show that there is still room for improvement.

1 Introduction

Visual Dialog (VD) is a task that combines natural language understanding grounded in vision with dialog. Being *visual*, VD is closely related to the area of Visual Question Answering (VQA). On VQA, important progress has been obtained recently with models that connect vision and language and are pre-trained on a variety of tasks (Tan and Bansal, 2019). Arguably, less progress has been made on the *dialog* part of VD, which is the topic of this paper. Currently, the two most popular datasets for visual dialog are VisDial (Das et al., 2017) and Guess-What?! (de Vries et al., 2017). The former contains chit-chat conversations about an image whereas the latter contains dialogs about a visual game whose goal is reference resolution, hence its dialogs are task-oriented. Reference resolution is a fundamental task in situated dialog (Clark and Wilkes-Gibbs, 1986; Clark, 1996; Foster et al., 2009). Questions in reference resolution can be classified as *intrinsic*

of the target (“It is a car?”) or *relative* to the context (“On the left?”) (Clark and Marshall, 1981).

Visual Dialog is assumed to require the dialog history to generate correct responses. However, it is not clear from previous work how dialog history is used for VD (Agarwal et al., 2020). In this paper we define history dependence in terms of a representation that is interpretable as a region of the visual common ground shared between dialog participants (Traum, 1994; Clark, 1996). This representation, which we call *Region under Discussion* (RuD), is inspired by the pragmatic theory of *Question under Discussion* (QuD). QuD (Roberts, 2012; Ginzburg, 2012; Velleman and Beaver, 2016) is a somewhat overlooked but conceptually fruitful theory for spelling out the connection between the information structure of a sentence or question and the discourse or dialog in which the utterance occurs. In this paper we define RuD and use it to connect a question to its visual dialog history; we make the following contributions:¹

- We define what it means for a visual question to require dialog history considering intrinsic and relative visual properties.
- We design a methodology for annotating a subset of the Guesswhat?! questions for which their dialog history is required because it completely changes their responses.
- We propose an interpretable representation of history based on the *Question under Discussion* (QuD) theory; we call our representation *Region under Discussion* (RuD).
- We extend the Oracle model by de Vries et al. (2017) and the LXMERT-based model of Testoni et al. (2020) with our RuD.
- We find that RuD summarizes dialog history in an interpretable visual way which is linguistically well founded and improves responses for history dependent questions.

¹Code and data at <https://github.com/mmazuecos/Region-under-discussion-for-visual-dialog>

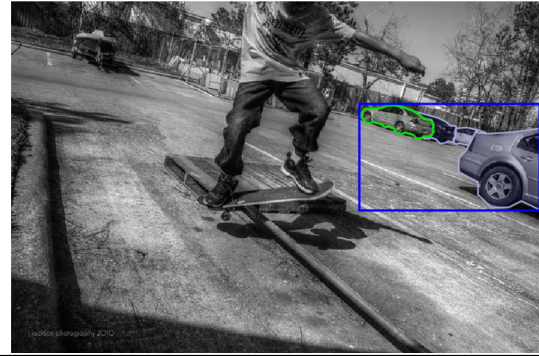
2 Region under Discussion (RuD)

Following Clark (1996) we define dialog *common ground* to be the commitments that the dialog partners have agreed upon during the dialog. An important part of the common ground is the *Question under Discussion (QuD)* (Ginzburg, 2012; De Kuthy et al., 2020). QuD is an analytic tool that has become popular among linguists and language philosophers as a way to characterize how a sentence fits in its context (Velleman and Beaver, 2016). The idea is that each sentence in discourse is interpreted with respect to a QuD. The QuD is defined by the dialog or discourse history. The linguistic form and the interpretation of an utterance, in turn, may depend on the QuD that provides the constraints that define the utterance’s context. Similarly, we define a *Region under Discussion (RuD)* for visual dialog as a representation of the constraints that the dialog history establishes. The interpretation of a question depends on its RuD.

Figure 1 shows a dialog from the GuessWhat?! visual dialog dataset (de Vries et al., 2017). Guess-What?! is a cooperative reference resolution game: two players attempt to identify an object in an image. The *Questioner* does not know the target object and has to find it by asking questions; the *Oracle* knows the target and provides yes/no answers. For each question in the dialog, its dialog history is defined as the previous questions together with their answers (DeVault et al., 2009). In the figure, the target is highlighted in green. The baseline Oracle model proposed by de Vries et al. correctly answers the first four questions, failing only in question number 5 with a *no* answer. This question does not look particularly difficult. So, why did it create a problem? *Because question 5 is the only question for which the dialog history modifies the response.* All the other questions can be answered correctly just by looking at the image and ignoring what was said before. That is, questions 1 to 4 are VQA turns because they do not need the dialog history.² If we answer question 5 *is it on the left?* ignoring the dialog history the correct answer is *no*, because the target is clearly to the right of the picture, not to the left. The RuD for this question, depicted in blue in the figure, modifies the response.

In this work, we model in the RuD the constraints that are related to intrinsic properties of the target that have been previously agreed upon

²We invite the reader to try it: just ignore the dialog history and answer the questions by only looking at the image.



Question	Human response
1. It is a person?	no
2. It is a car?	yes
3. Is it in the back?	yes
4. Are there two together?	yes
5. <i>Is it on the left?</i>	yes

Figure 1: Human-human dialog from the GuessWhat?! dataset (de Vries et al., 2017). The example illustrates our definition of history dependent question. Question 5 can be correctly answered with *no* if asked at the beginning of the dialog, when the dialog history is empty because the target (marked in green) is not to the left of the picture. However, when the RuD (depicted in blue) is constrained by the initial turns then the correct answer to the same question is *yes*.

between the dialog participants. An *intrinsic* property is one that is inherent and inseparable from the target and is not dependent on the visual context that the target is put in. In this example, such intrinsic property is the fact that the target is a car, which is established in question 2. Another intrinsic property may be that the target is a vehicle, but not the fact that the target is together with another car. We say that such property is not intrinsic of the target but relative to the position of the car. We decide to represent in the RuD only intrinsic history motivated by literature from robot dialog, where intrinsic properties are plentiful and stable constraints (Tan et al., 2020). Using intrinsic properties appears as the most common strategy for recovering from ambiguous dialog situations, as they reduce the cognitive effort (Marge and Rudnicky, 2015). We believe that restricting the RuD to intrinsic properties allows us to focus on the phenomena we are interested in while keeping the model simple and easily interpretable.

Summing up, most questions in this dialog *can* be correctly answered independently of the dialog: they do not need the history. In effect, except for

Type	Quantity	Sample question
Object	39269	<i>is it a traffic light?</i>
Spatial	39250	<i>is it on the left?</i>
Color	15403	<i>its color is light brown?</i>
Other	7925	<i>do you sit on it?</i>
Action	7645	<i>is he running?</i>
Size	1364	<i>the big one?</i>
Texture	901	<i>a rough surface?</i>
Shape	301	<i>the round one?</i>

Table 1: Question type distribution in successful games in the test set (de Vries et al., 2017), following the classification proposed in (Shekhar et al., 2019).

one turn, Figure 1 is just visual question answering. In this paper we model dialog history as constraints that represent the part of the image which the dialog partners agree is the RuD and over which the rest of the questions are to be interpreted. For our example, with respect to the blue box, the correct answer of *Is it on the left?* is yes since the car is on the left of the agreed RuD.

3 Methodology

In this section we describe the dataset and we show how we annotate a subset of questions whose dialog history completely change their responses. We then explain how we build a semantic history for each dialog in order to construct a RuD and how we extend Oracle models with RuDs.

3.1 Dataset and annotation

The GuessWhat?! dataset (de Vries et al., 2017) contains around 135k successful human-human dialogs with an average of 5 questions in natural language created by crowdsourcers playing the reference game on MS COCO images (Lin et al., 2014). The set contains around 672K questions which are grounded on about 63K unique images. Following Shekhar et al. (2019), we classify the questions into different types. In Table 1, we show the test set support for each type as well as a sample question.

The table shows that the most frequent types of questions in the dataset are object and spatial questions. They constitute about 40% of the total questions. Object questions are intrinsic and do not depend on the RuD to be interpreted. Differently, spatial, color and size questions are relative and can have their meaning changed due to the RuD as defined and illustrated in Sections 2 and 4.

To spot history dependent questions, we first sample a set of relative questions that follow a positively answered object question in a dialog. Then, two annotators identify questions such that the polarity of the answer changes when the question is asked considering its history. The annotation procedure is as follows: (1) Look at the picture and the candidate question without looking at the dialog history. (2) Answer the question with “yes”, “maybe yes”, “maybe no”, “no” or “I don’t know” (3) Compare to the answer in the corpus that the person gave to that question considering the dialog history. (4) If the answers do not coincide, mark the question as history dependent.³ In this setting, disagreements between annotators mostly arise from different views on vague properties of objects.

Surprisingly, and in contrast to what is usually assumed in previous work (Agarwal et al., 2020), visual questions dependent on dialog history do not contain more pronouns and ellipses than history independent visual questions. From the 1658 questions analyzed, two annotators agreed that 204 questions are history dependent. We call these 204 questions our GWHist test set⁴. By this procedure, we marked 12.3% of the questions in the sample as history dependent.

3.2 Semantic history

To build the RuDs, we parse and match the questions in each dialog history to build a *semantic history*, this is, a representation of the known intrinsic properties of the target object. Then, we use this information to filter the objects in the image and obtain a set of *candidate objects* that will be part of the RuD.

Parsing. We parse questions that establish relations of types “*is a*” and “*is the*” between a noun phrase (NP) and the target object. The answers to these questions usually convey information about the category of the object, as in “Is it a person?”. A positive answer to a *category question* implies that the candidate objects include only objects of that category, while a negative answer implies that these objects are not candidates.

We define regular expressions for the most common syntactic patterns. We tokenize and POS tag the questions using NLTK and Stanza (Bird et al.,

³See Appendix C for a screenshot of our annotation tool.

⁴The class balance of the GW test set is: 50.5% are answered with “No”, 47.7% with “Yes” and 1.8% with “N/A” (Non Answerable). In our GWHist we exclude the “N/A” class, “Yes” is 50% and the “No” class the other 50%.

Pattern	Example
NP?	1. <i>person</i> ?
is it a NP ?	2. <i>is it a red car</i> ?
is the NOUN a NP ?	3. <i>is the object a plate</i> ?
is it one of the NP ?	4. <i>is it one of the boats</i> ?
NP = NOUN NOUN NOUN ADJ NOUN	

Table 2: Common syntactic patterns for category questions and examples for them. The patterns for NPs are defined at the bottom.

2009; Qi et al., 2020). Table 2 shows some of the main patterns we use.

Matching. After parsing, the obtained NPs are lemmatized using NLTK and matched to the 80 categories from the COCO dataset. Lemmatization is particularly useful to match questions using plural nouns (as “boats” in example 4, Table 2). Matching is done using exact string comparison. Two complementary matching strategies are discussed in the following two paragraphs.

In the case of some category questions with two-token NPs, only the second token refers to the category, while the first one refers to another intrinsic property (as color in example 2, Table 2). In this case, we match only the second token to a category, if the answer is positive. A negative answer is not informative about the category (it may be a green car).

Some NPs refer to categories not present in COCO but to *supercategories*, i.e., nouns that cover several COCO categories (e.g. “food”, covering “apple”, “banana”, “broccoli”, etc.). We match these nouns using a pre-computed list of known supercategories. The supercategories, and its mapping to categories, are obtained from WordNet (Fellbaum, 1998) by extracting hypernym relations.

Filtering. The parsing and matching processes result in a *semantic history* that is available for each question in a game. The semantic history is the ordered list of positive and negative relations to (super)categories found in the previous turns (e.g. [(*pos*, “vehicle”), (*neg*, “car”)] means that the target is a vehicle but it is not a car). The objects in the image are filtered using the history to obtain a set of candidate objects. Next, we describe our approaches for positive and negative elements of the history separately.

For the positive history we use only the last el-

ement, assuming that it is the most specific one. We select the objects that are consistent with the (super)category of this element. For the negative history, our policy is to remove all the objects in the negated (super)categories from the candidates. For example, in Figure 1 the RuD after question 1 is answered with *no* removes the boy on the skateboard from the candidates. Here, we assume that all the negative elements identify objects that can be removed from the RuD, regardless of the order in which they appear.

After processing the semantic history, we check the candidate objects set for well-formedness. We say that the set is *ill-formed* if it does not include the target object. In this case, we force the inclusion of the target object as an *ad-hoc* policy.

Coverage. To evaluate the coverage of the semantic history, we apply it to the validation set of the GuessWhat! dataset. In addition to the full-featured process, we try three feature ablations by removing either supercategory matching (**-super**), second-token matching (**-2nd**) or negative history (**-neg**). This way, we are able to assess the individual contribution of each of these features.

A summary of the coverage is shown in Table 3. We report here the total number of questions with non-empty semantic histories, and the counts for different types of candidate objects sets: ill-formed sets such as empty ones (**empty**) and those that exclude the target (**w/o tgt**), and well-formed sets such as those that only include the target (**only tgt**) and those that include the target and some other distractor objects (**tgt+dist**).

Despite the simplicity of our approach, there is an important coverage of the questions, with more than 60% having semantic history. We also see that there is a low rate of ill-formed candidate sets of $\sim 3\%$.⁵ Ablations show that, as expected, negative history almost doubles the coverage. Also, WordNet-based supercategories makes an important contribution to coverage, at the expense of a significant increase on ill-formed candidate sets.

3.3 Extending oracle models with RuD

In this section we extend two popular models for the Oracle in visual dialog, namely the Question+Category+Spatial (QCS) baseline proposed by de Vries et al. (2017) and the more recent LXMERT-based cross-modal Oracle (CMO) pro-

⁵Recall that ill-formed sets are fixed by forcing the addition of the target to them.

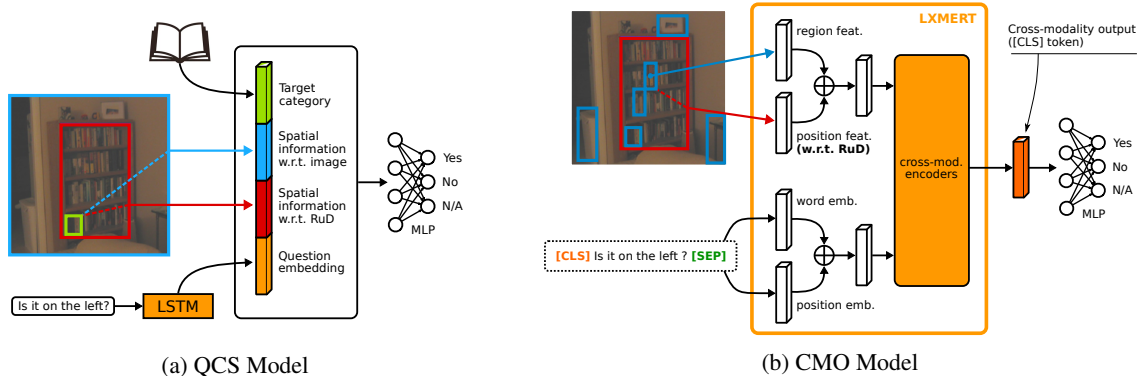


Figure 2: Architectures and inputs for both QCS and CMO models extended with our RuD representation.

Questions	full	-super	-2nd	-neg
Has hist.	62096	54019	61743	34151
%	63.0%	54.8%	62.7%	34.7%
candidates	empty	1529	507	1386
	w/o tgt	894	401	841
	only tgt	3444	2681	3415
	tgt+dist	56229	50430	56101

Table 3: Number of questions with history for the 98507 questions in the validation dataset. Also, detail for different kinds of candidate objects sets.

posed by Testoni et al. (2020). QCS was shown to be the best performing baseline in de Vries et al. (2017) and has become the most frequently used Oracle (de Vries et al., 2017; Strub et al., 2017; Shekhar et al., 2019; Pang and Wang, 2020). CMO improves over the QCS baseline by taking advantage of the powerful multi-modal LXMERT encoder (Tan and Bansal, 2019), showing SOTA performance for the task. In what follows, we build upon these models and propose two simple extensions to encode the RuD. We name our models as QCS+RuD and CMO+RuD, respectively.

For both models, we define the RuD as the smallest bounding box that encloses all the objects in the set of candidates. The candidates objects are computed from the dialog history as described in 3.2. If no history is available we set the RuD to match the whole image.

QCS takes as input a question encoded by an LSTM as well as category and spatial feature embeddings of the target. An MLP on top of these features classifies the question into three possible answers: *no*, *yes* and *n/a* (*non answerable*). The spatial embedding in QCS corresponds to an 8-dimensional vector that encodes the coordinates of the top-left and bottom-right corners, center

and size of the target bounding box, normalized such that the image width and height coordinates range from -1 to 1. We extend this encoding by adding the same 8-dimensional vector but shifted and scaled according to the RuD position and scale. Concretely, let (x_1, y_1, x_2, y_2) be top-left and right-bottom coordinates of the target bounding and (X_1, Y_1, X_2, Y_2) that of the RuD. Let us define $x_0 = (x_1 + x_2)/2$, $y_0 = (y_1 + y_2)/2$ and let (w, h) and (W, H) denote the width and height of the target box and RuD, respectively. We add the following features to the QCS input embedding: $2 \frac{x_1 - X_1}{W} - 1$, $2 \frac{y_1 - Y_1}{H} - 1$, $2 \frac{x_2 - X_1}{W} - 1$, $2 \frac{y_2 - Y_1}{H} - 1$, $\frac{x_0}{W}$, $\frac{y_0}{H}$, $\frac{w}{W}$ and $\frac{h}{H}$. The proposed architecture is shown in Figure 2a.

For questions without history, the RuD spatial embedding is defined to be the same as the spatial embedding w.r.t. the entire image.

For CMO, the model expects as inputs not only word and region embeddings but also their location with respect to the query and reference image, respectively. For the visual modality, this information is encoded in the form of bounding box coordinates after the object detection module. In our case, this corresponds to the coordinates of the top-left and bottom-right corners of each object bounding box. Using the same notation as before, we encode each box spatial coordinates as $(\frac{x_1 - X_1}{W}, \frac{y_1 - Y_1}{H}, \frac{x_2 - X_1}{W}, \frac{y_2 - Y_1}{H})$. In Figure 2b we show how we implement RuD for CMO. Note that, in this case, coordinates lying outside the RuD will be negative or with a value greater than one. This does not happen for the QCS+RuD model because only the coordinates of the target are modified and these always fall inside the RuD.

4 Results and discussion

In this section we first report the empirical results of our experiments, then we argue that RuD summarizes history in a visually interpretable way through a qualitative analysis. Finally we discuss the limitations of our implementation of RuD.

We performed our experiments with the previously proposed models for the Oracle task. We implement both of our models as three-way classifiers using MLPs and a cross-entropy loss, accordingly with the relevant literature. For the QCS baseline, we follow de Vries et al. (2017) and use a two layer MLP with ReLU non-linearities (1024-ReLU-128) while for the LXMERT-based Oracle we use a simpler setup with just one layer on top of the cross-modality output of LXMERT. Our CMO implementation is based on the pre-trained LXMERT model from the Transformers library (Wolf et al., 2020). Visual features are the same as in Testoni et al. (2020). We leave the rest of the details of our experiments in the Appendix A.

4.1 Empirical results

We report empirical results for the Oracle task of the GuessWhat?! benchmark (de Vries et al., 2017) and for the history dependent subset GWHist described in Section 3. We evaluate the RuD-augmented models and compare them with their respective RuD-less baselines.

In Table 4 we show the accuracy in the test set of each of our models for the questions that were augmented with a semantic history. We use Oracle response accuracy as an evaluation metric because it compares the model response to the human ground truth answer. In addition to the GuessWhat?! test set and our GWHist subset, we report the results on the two more frequent types of questions: object and spatial⁶. The table shows that the RuD-augmented models do not outperform the RuD-less models on the object subset. This is to be expected, since object questions are not history dependent. We anticipated this in Section 2. For example, a car will always be a car no matter what was said about it before.

The accuracy for spatial questions and the whole GuessWhat?! (GW) dataset is slightly higher for the models that add RuD but the difference is small. This is due to the fact that most questions in GW including spatial questions are not history dependent

⁶The analysis of the accuracy across all types of questions is included in Appendix A.

Type	QCS	QCS+RuD	CMO	CMO+RuD
Object	0.901	0.902	0.894	0.894
Spatial	0.669	0.691	0.770	0.777
GW	0.733	0.744	0.809	0.813
GWHist	0.285	0.402	0.285	0.416

Table 4: Test response accuracy for the Oracle models discussed in Section 3 with and without Region under Discussion (RuD). Results are shown for the question types object and spatial. Last two rows show the accuracy on the whole test set (GW) and on a history dependent subset (GWHist)

as we argued in Section 3. However, the effect of adding the RuD on accuracy is clear in the history-dependent GWHist, where QCS+RuD and CMO+RuD show an increment of 41% and 46%, respectively. The initial accuracy for the GWHist is very low for both QCS and CMO models. In fact, the accuracy is close to one minus the accuracy on GW. These are hard questions that are wrongly answered without the dialog history as we explained in Section 3. The fact that both models consistently improve shows that the RuD is capturing the region of the image on which the history dependent question is being interpreted. With a 0.416 maximum accuracy for history dependent questions there is still a lot of phenomena that our models are not able to handle. Below we discuss the kinds of history dependent questions that our models are able to handle and also illustrate those that they cannot.

4.2 Qualitative analysis

In this subsection we argue that RuD summarizes history in an interpretable visual way for different types of questions. Size, color and spatial questions can have a meaning which is relative to their RuDs. We also discuss details about the GWHist and we show examples of the phenomena we found during the annotation.

In the first example we see a size question, *the big one near the white plate?* in position 8, that gets correctly answered by the RuD-augmented CMO. In this picture, the target is the biggest bottle visible marked in green. The model can use the RuD to determine which of the biggest bottle relative to the other bottle present on the scene.

The second image shows an example of a color questions that improves when answered within the RuD. The model is able to take advantage of the RuD to answer the question *it is brown?* on position 4. Despite the car being some gamma of gray in




	Question	HR	CMO	+RuD
	1. is it human?	no	no	no
	2. is it food?	no	no	no
	3. is it on the gas stove?	no	no	no
	4. is it on the nearby counter top?	yes	yes	yes
	5. is it red?	no	no	no
	6. is the yellow spoon in the plate?	no	no	no
	7. is a bottle?	yes	no	no
	8. <i>the big one near the white plate?</i>	yes	no	yes
	1. it is a sign?	no	no	no
	2. it is a car?	yes	yes	yes
	3. it is grey?	no	no	no
	4. <i>it is brown?</i>	yes	no	yes
	5. it is front the other car?	yes	no	no
	1. is it a vehicle?	no	no	no
	2. is it a person?	no	no	no
	3. is it a building?	no	no	no
	4. is the color red?	no	no	no
	5. is it the sign board?	no	no	no
	6. is it a traffic light?	yes	yes	yes
	7. is it in middle?	no	no	no
	8. <i>is it the first one?</i>	yes	no	yes

Figure 3: The questions in italics are history-dependent. They illustrate how different kinds of questions may need to be interpreted respect to the RuD. CMO does not answer these questions correctly, but CMO+RuD does. The RuDs are in blue. The targets are in green. HR is the human response. The questions in italics from top to bottom are size, color, and a kind of spatial question that specifies order.

the illumination conditions of the scene and given the answer “no” to the question *it is grey?* before, we could make an argument that the target is the *browner* object in the region.

In the third example, question 8 *is it the first one?* is interpreted with respect to a thin and long RuD which establishes an order in the traffic lights.

In Figure 4 we show an example of a history dependent question that is not improved with the RuD-augmented models. In this case, the question is *most first?*. This example shows one of the limitations of our approach. A model that would correctly answer these sorts of questions would need to take into account the second question *in right?* to infer the direction of the search and arrange candidate objects in a row indexed from right to left.

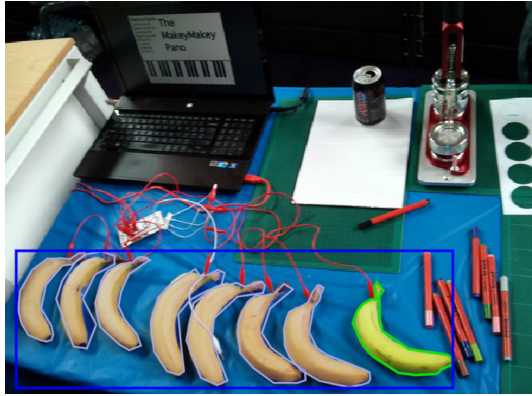
During annotation we also found a variety of examples of questions that asked for objects other than the target. These questions change their se-

mantics completely when isolated from the dialog history. We found that many of these history dependent questions come from an object question that has already identified the category of the target object and now are looking for another salient object to univocally identifying it. We show examples of this and other history dependent questions that our models are not able to handle in the Appendix B.

Additionally, the GuessWhat?! dataset was generated by crowdworkers and some of the questions exhibit English errors. An example of this can be seen in the third question in Figure 4.

4.3 Limitations

In this work, we relied on the annotations of the COCO dataset to compute the RuDs. However, dialogs may contain questions that refer to objects not present in the annotations; those objects are invisible to our RuD computation. Depending on the COCO annotations makes it easy to compute RuDs



Question	Human response
1. is a banana?	yes
2. in right?	yes
3. <i>most first?</i>	yes

Figure 4: Example of a GWHist example that does not improve with our approach. Such an example is hard as it will need further dialog management to get that the questioner’s attended point is at the right of the bananas and that a row of such would be indexed from right to left.

with intrinsic history. The same cannot be said about histories regarding attributes such as color, size, shape, etc. Dialogs contain questions that rely on these questions to build common ground.

Lastly, many Questioners further constrain the RuD multiple times (either by using grouping, filtering by attributes, delimiting the area with respect to another object, etc). This process requires more history management than we do to compute the RuD for a given question. Most of these constraints require common sense reasoning, spatial understanding and a deep connection to the visual modality. As we explained in Section 2 in this paper we only consider intrinsic properties (that is, object questions) to constrain the RuD. This approach is not enough, for example, if the question 5 in Figure 1 would have been “*Is it on the right?*” the RuD would be too large.

5 Previous work

Visual Dialog played a prominent role in early work on natural language understanding (Winograd, 1972) and is now the focus of an active community investigating the interplay between computer vision and computational linguistics (Baldrige et al., 2018; Shekhar et al., 2019). On the GuessWhat?! task, most previous research has focused on the

Questioner (Strub et al., 2017; Shekhar et al., 2019; Pang and Wang, 2020). Recent work suggests that the performance of the *Oracle* agent used by most work (de Vries et al., 2017) is quite different for types of questions (Mazuecos et al., 2020). Questioners that rely on the Oracle learn to prefer to ask only those questions that the Oracle can answer reliably. This has an impact on the type and linguistic variety of the generated questions, reducing the Guesswhat?! task to a simpler linguistic task (Shukla et al., 2019; Pang and Wang, 2020).

Clark and Wilkes-Gibbs (1986) models the process of finding referring expressions as a collaborative process in which the speakers repair, expand on, or replace the noun phrase in an iterative process until they reach a version they mutually accept. This process is explicitly performed in a Guesswhat?! dialog although the role of the Oracle is simplified.

The Oracle model proposed by de Vries et al. (2017) is implemented with an MLP (as we described in Section 3). They showed that their best performing model was the one that takes the question, the target’s category and its location as inputs. This has a major limitation: the model is blind and cannot see the image. This proposed model is widely used as the *Oracle* agent for all of the following research on the *Questioner*.

Testoni et al. (2020) proposed an adaptation of LXMERT (Tan and Bansal, 2019) to improve on the previous Oracle, achieving a new SOTA for the GuessWhat?! Oracle without using dialog history as an input. This work showed various improvements in different types of questions, mainly on questions regarding location and other attributes and a little decrease in performance on object or super category questions due to not receiving the gold standard object category as input from the dataset. Their qualitative error analysis suggests that spatial questions are harder because they require history in order to be answered correctly in context.

Agarwal et al. (2020) argues that although complex models that encode history for visual dialog have been proposed (Yang et al., 2019), such work has not demonstrated that history indeed matters for visual dialog. Agarwal et al. propose and apply a new methodology for evaluating history dependence of questions in visual dialog. They show crowdsourcers a question with its image without the dialog history and ask the crowdsourcer “would you be able to answer this question by looking at

the image only or you need more information from the previous conversation?”. But saying *I can confidently tell the correct answer just by looking at the image* is not the same as answering it in the same way that one would by looking at the previous conversation (remember the example in Section 2). Most questions are answerable no matter where they appear in a dialog because the answerer accommodates. Our method differs in that our’s has the advantage of getting history dependent questions that are not evident at first glance (such as “*is it on the left?*” in Figure 1). We found a similar percent of questions in the GW dataset that are history dependent, as Agarwal et al. did on VisDial (12% vs. 11%). This may result in current dialogue models not learning history dependence since current mainstream vision and dialog datasets lack a significant amount of history dependency.

Dialog history has *two* characteristics that makes it difficult for current machine learning methods: not only it introduces variability with different histories for the same question, history dependence may also not be lexicalized, as in *is it on the left?* in Figure 1. History dependency is easier to spot when it is lexicalized with explicit pronouns (e.g. *him* in ‘is it close to him?’) or through noticeable ellipsis (e.g. a missing noun such as *cars* in ‘are there two together?’). However, as we see in Figure 3, pronouns in task-oriented VD frequently are not anaphoric to the dialog history but to the image (e.g. the pronoun *it* in *is it a person?* is anaphoric to the target). Information structure theory (Roberts, 2012) and, in particular, QuD (Purver et al., 2003; Ginzburg, 2012; De Kuthy et al., 2020) provide a framework for defining context dependence beyond pronouns and syntactic ellipsis.

6 Conclusions

We proposed a novel *interpretable* representation for visual dialog history: Region under Discussion (RuD). It constrains the image spatial features according to a semantic representation of the history inspired in the information structure notion of QuD. We evaluated our method on models for the Oracle task in the GuessWhat?! dataset. Our results show that our implementation of RuD leads to improvements in performance on history dependent questions. We release a manually annotated subset of such questions. Our experiments confirm that intrinsic properties do not benefit from dialog management whereas questions that ask for properties

relative to the context see an improvement with it.

Interestingly, only a low percentage of questions (12%) are indeed history dependent in the Guesswhat?! dataset. However, a single error in a 10 turns GW dialog may cause the identification of the wrong referent, rendering the task unsuccessful. We agree with de Vries et al. (2020) that the simplified yes-no nature of this task allows us to focus on an interesting playground for working on conceptual advances in representation methods for dialog history. The Guesswhat?! task is ill-suited for incremental research, as it is unclear how small improvements will find their way to real applications. Our contribution is not incremental. Our paper makes a theoretical contribution by defining the new concept of Region under Discussion and linking it with the concept of Question under Discussion in dialog. Based on this theoretical contribution it proposes an interpretable, simple and extensible method for representing dialog history.

This work only adjusts the RuD to reflect the intrinsic properties of the target entity, not other attributes (color, shape, etc.) and spatial restrictions (“is it among the four in the back?”). Including other types of relations in the generation of the RuDs is a promising avenue for future research. In this regard, we are considering the following approaches: 1) RuD generation from scene graphs (a SG is a graphical representation of an image that encodes objects as nodes and pairwise relations as edges), and 2) learning RuD predictors from dialog data end-to-end. In both cases, we need a large and representative training set (SG/RuD annotated for each turn on each dialog) and such data is hard and expensive to gather. A possible solution in this case is to explore weakly supervised strategies, where the SG/RuD is treated as a latent variable.

We think that these contributions can be of use for the Questioner model, potentially helping Questioners learn dialog strategies instead of solving dialog tasks through Visual Question Answering.

Acknowledgements

We are grateful to Patrick Blackburn and Sasha Luccioni for their helpful comments. This work was supported by SECyT, Universidad Nacional de Córdoba (projects 32720200400442CB and 33620180100076CB), and used computational resources from CCAD-UNC, which is part of SNCAD-MinCyT, Argentina.

Ethical considerations

In this paper we trained simple and complex deep learning models. We have consumed approximately 16.5Wh for each experiment with QCS and 266.67Wh for each one with CMO. We generated approximately 0.02 kgCO₂eq and 0.77 kgCO₂eq for each QCS and CMO experiment, respectively⁷. Each QCS experiment took approximately 9min to train its 4.3M trainable parameters. It raises to around 6.7hs to train the 207.94M parameters of the CMO models. We have not collected a new dataset so we have not used crowdsourcing. The annotation of the GWHist corpus was done by two of the authors who were not economically rewarded. However, this work builds upon work or which carbon footprint and the ethical considerations of crowdsourcing are important. We discuss these ethical considerations below.

First, the dataset that we use in this paper is described in (de Vries et al., 2017) which was crowdsourced. Crowdsourcing raises ethical concerns including paying a fair wage to crowdworkers, and limiting the amount of hits they make in a day so that they are not exhausted and overworked. de Vries et al. (2017) do not provide this information in their paper. Last but not least, machine learning models trained on long multimodal dialog histories may get very big very fast (Agarwal et al., 2020). We need models that learn to summarize dialog histories as we do with RuDs for the sake of the environment and the budget of low-resource researchers.

References

- Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.
- Jason Baldridge, Tania Bedrax-Weiss, Daphne Luong, Sridhar Narayanan, Bo Pang, Fernando Pereira, Radu Soricut, Michael Tseng, and Yuan Zhang. 2018. Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *Proceedings of the First International Workshop on Spatial Language Understanding*, New Orleans. Association for Computational Linguistics.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly, Beijing.
- Herbert Clark. 1996. *Using Language*. Cambridge University Press, New York.
- Herbert Clark and Catherine Marshall. 1981. Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Elements of discourse understanding*, pages 10–63. Cambridge University Press.
- Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.
- Kordula De Kuthy, Madeeswaran Kannan, Haemant Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798.
- Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv*, abs/2007.14435.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4475. IEEE Computer Society.
- David DeVault, Kenji Sagae, and David Traum. 2009. Can I finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue*. Association for Computational Linguistics.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.
- Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California.
- Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford Press.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision*

⁷A 60W lamp on for 24hs generates 0.67 kgCO₂eq according to the local emission factor.

- *ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer.
- Matthew Marge and Alexander Rudnicky. 2015. Miscommunication recovery in physically situated dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 22–31, Prague, Czech Republic. Association for Computational Linguistics.
- Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti. 2020. On the role of effective and referring questions in GuessWhat?! In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 19–25, Online. Association for Computational Linguistics.
- Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11831–11838. AAAI Press.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and New Directions in Discourse and Dialogue*, pages 235–255. Kluwer Academic Publishers.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and Guess-What. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2578–2587. Association for Computational Linguistics.
- Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should I ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451, Florence, Italy. Association for Computational Linguistics.
- Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of international joint conference on artificial intelligence (IJCAI)*.
- Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.
- Xiang Zhi Tan, Sean Andrist, Dan Bohus, and Eric Horvitz. 2020. Now, over here: Leveraging extended attentional capabilities in human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI ’20*, page 468–470, New York, NY, USA. Association for Computing Machinery.
- Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. 2020. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 29–38, Online. Association for Computational Linguistics.
- David Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Computer Science Dept., U. Rochester, USA. Supervised by James Allen.
- Leah Velleman and David Beaver. 2016. Question-based models of information structure. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*. Oxford University Press.
- Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3:1–191.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2561–2569. IEEE.

A Towards reproducibility

A.1 Implementation details and architecture

Learning rate, epochs and batch size are set to $\{10^{-4}, 16, 1024\}$ and $\{10^{-5}, 5, 32\}$ for the QCS and CMO oracles, respectively. We ran a grid search only for the CMO baseline over the range $\{3e-4, 1e-4, 1e-5, 1e-6\} \times \{16, 32, 64, 128\}$ and 5 training epochs. We choose the best combination by monitoring accuracy on the validation set. We implemented our models in PyTorch (Paszke et al., 2019) and trained them on NVIDIA GTX 1080 Ti GPUs.

A.2 Extended Empirical results

In this subsection we show the extended results displayed by games with and without history and by type of question.

Type	QCS	QCS+RuD	-super	-2nd	-neg
object	0.913	0.911	0.910	0.911	0.910
spatial	0.677	0.695	0.691	0.693	0.693
color	0.627	0.632	0.629	0.630	0.631
action	0.652	0.651	0.649	0.647	0.647
size	0.639	0.672	0.661	0.655	0.661
texture	0.716	0.722	0.715	0.716	0.701
shape	0.720	0.700	0.713	0.703	0.713
GW	0.786	0.792	0.790	0.791	0.791

Table 5: Results in the validation set for the QCS Oracle models. Total classification accuracy and accuracies for the different question types are reported. The QCS baseline is compared to our QCS+RuD models and three feature ablations of QCS+RuD.

QCS and QCS+RuD. Table 5 show results for QCS and QCS+RuD. Results are shown for the full featured RuD, as well as for the three feature ablations discussed in Section 3. The RuD-augmented models outperform the QCS baseline for questions that can be considered relative: spatial, size, and

(arguably) color, texture and action. No improvement is observed for intrinsic questions: object and shape. Spatial questions, the most frequent relative question type, are the most benefited by the use of the RuD, with improvements in accuracy ranging from 1.4% to 1.8%. Ablations show that all the proposed features contribute to the overall performance, with the use of WordNet-based supercategory being the most contributing one.

We experimented with word embedding to retrieve the semantic histories instead of using the semantic parser we proposed in Section 3. The relations between the content of a sentence and the generated histories were calculated using cosine distance. For that we tried different thresholds. A manual analysis showed that higher thresholds let too many errors in while lower thresholds got lower coverage than the proposed method. We then decided to stick with our semantic parser and leave the exploration of word embeddings for future work.

These empirical results suggest that our RuD seems to be capturing a fact about language: relative questions tend to depend on dialog history while intrinsic questions do not. However, the improvement on relative questions is small. We believe that working on more elaborate semantic history and RuD construction schemes can lead to further significant improvements in Oracle performance.

Final results for the test set are shown in Table 6. Accuracies are reported for the different question types, and also for questions with and without RuD (“w” and “w/o” resp.).

CMO and CMO+RuD. Results are shown in the last two groups of columns in Table 6. Compared to the QCS counterparts, we see an increase on performance for all question types, consistent with Testoni et al. (2020) results. Accuracy improves on more than 5 absolute points for CMO and CMO+RuD compared to QCS and QCS+RuD, respectively. However, when considering CMO vs. CMO+RuD we observe only marginal improvements on spatial, color, size and action questions. Consistently with what was found for the QCS+RuD model, intrinsic questions object and shape do not show improvements for CMO+RuD. The only significant improvement is observed in the spatial subset of history dependent questions (GWHist). Here we observe a large gap in favor of the CMO+RuD model on questions with RuD.

Type	QCS			QCS+RuD			CMO			CMO+RuD		
	all	w	w/o	all	w	w/o	all	w	w/o	all	w	w/o
object	0.936	0.901	0.967	0.936	0.902	0.965	0.923	0.894	0.947	0.922	0.894	0.946
spatial	0.675	0.669	0.697	0.694	0.691	0.702	0.775	0.770	0.792	0.780	0.777	0.790
color	0.615	0.604	0.651	0.623	0.615	0.651	0.777	0.769	0.805	0.778	0.772	0.800
action	0.641	0.620	0.718	0.650	0.628	0.729	0.780	0.769	0.820	0.785	0.776	0.820
size	0.614	0.595	0.678	0.639	0.630	0.671	0.751	0.748	0.763	0.757	0.756	0.763
texture	0.719	0.716	0.725	0.704	0.688	0.731	0.789	0.775	0.814	0.788	0.782	0.798
shape	0.674	0.659	0.695	0.678	0.659	0.703	0.751	0.740	0.766	0.757	0.734	0.789
GW	0.782	0.733	0.864	0.788	0.744	0.863	0.839	0.809	0.891	0.841	0.813	0.889
GWHist	0.299	0.285	0.333	0.392	0.403	0.366	0.259	0.285	0.200	0.357	0.417	0.216

Table 6: Test classification accuracy for the Oracle models discussed in Section 3. Results are shown for different question types and for questions with and without history information (RuD). Last two rows show the accuracy on the whole test set (GW) and on a history dependent subset (GWHist).

We also consider a control configuration based on zeroing the spatial information associated to the visual input modality. This model obtains an overall accuracy of 0.750 on the full test set, a value that is below that obtained with the QCS baseline. This shows the importance of the spatial information for these types of models. When we consider the performance of this model on the GWHist subset, performance is around 50% (0.515, 0.566 and 0.300 for “all”, “w” and “w/o”, respectively). This is to be expected since the GWHist subset was designed such that the absence of history information would change the polarity of the answer. The observed 50% is close to a history-less majority class predictor. The performance of CMO+RuD of 0.301 on the history dependent set GWHist leaves much room for improvement. Below we illustrate the performance of CMO+RuD for relative questions. Then we turn to the limitations of our approach.

B More Qualitative Analysis

As explained in the qualitative analysis on Section 4, there were some questions that some questions asked for objects other than the target. An example of this can be seen in the first image in Figure 5. In isolation, final question “*guy in red?*” would be interpreted as just another object question, but in the context they are trying to identify another object in the image that is related to the target. This relation is usually set on before such question appear.

Some more limitations present in the data are that, sometimes, referenced objects are not present in the annotation. The second example in Figure 5 shows this phenomena. The fourth question, “*is it*

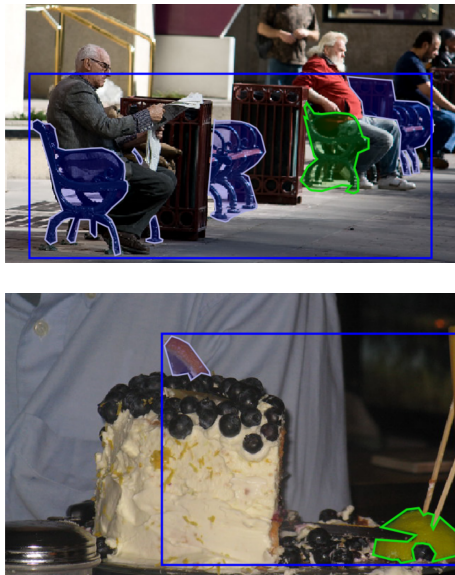
small?” has the correct RuD, but fails to compare the orange to the little blueberries when answering to the question.

Coming back to the spatial questions, questions that ask for absolute spatial location of objects tend to have more ellipsis than other questions. Non history-dependent absolute spatial questions do not differ syntactically from their history dependent counterparts. It is when analyzed in context that one can start making distinctions between one and the other, but that escapes the form of the sentences and requires knowledge that the RuD-less models did not have access to before.

Exophora In the Guesswhat dataset we find that visual questions dependent on dialog history do not contain more pronouns and ellipses than history-independent visual questions, as said in section 3. This is due to most questions having exophora in the corpus, relying heavily on the common visual context. Such exophoric pronouns are grounded in the task and the image and not in the previous dialogue. Exophoras not only refer to the target, but also to other salient objects that can be referred to with a pronoun without being linguistically introduced. For example, in a picture with 2 salient people a question such as “*is it behind them?*” is possible when the people were not referred to before.

C Annotation Tool

We used a web interface as shown Figure 6 for the annotation of the data. Each annotator was prompted with a question and were asked to answer the question with one of the 5 options shown in the



Question	HR	CMO	+RuD
1. is it a person?	no	no	no
2. is it a bench?	yes	yes	yes
3. is it the leftmost bench?	no	no	no
4. the second bench from left?	no	no	no
5. is there someone sitting on it?	yes	yes	no
5. <i>guy in red?</i>	yes	no	yes
1. is it any type of food?	yes	yes	yes
2. is it a fruit?	yes	yes	yes
3. is it shaped like a ball?	yes	yes	yes
4. <i>is it small?</i>	no	yes	yes
5. is it orange?	yes	yes	yes
6. is it tangy?	yes	yes	yes

Figure 5: Examples of the GWHist questions (marked in italics) that exhibit complex phenomena (such as questions that ask for objects other than the target) and the lack of annotations in the GuessWhat?! dataset. All RuD-augmented models fail on the history dependent questions.

figure: “No”, “Maybe no”, “I don’t know”, “Maybe yes”, “Yes”. Once the answer is decided, the system would log the annotation and prompt the user with a new pair of question-image.

For annotation we loaded 11306 color, size and spatial questions from the GuessWhat?! sampled from the set of questions that follow an object questions. The pool was formed by:

- 4141 color questions
- 431 size questions
- 6704 spatial questions

We assigned different pool of questions to each pair of annotators as for them to cover as much as possible from the questions’ pool and to make sure each annotated question had at least 2 annotators. From that pool we analyzed the 1658 questions referenced in Section 3. We ended up with a GWHist dataset that contained 204 history dependent questions whose question types where distributed as follows:

- 2 action questions
- 192 spatial questions
- 23 color questions
- 6 size questions
- 1 texture question

Keep in mind that a question can fall in multiple of these question types. For example “*the one on the right that has a little red left in it?*” (present in the GWHist) is classified as a color, size and spatial question and, as such, is counted for each of the

question types.

The questions were mapped to their corresponding answer given in the GuessWhat?! convention: answers “yes” and “no” are kept the same, “I don’t know” questions are mapped to “N/A”. “Maybe yes” and “Maybe no” are mapped to “yes” and “no”, respectively. Once the annotator answers are mapped we can compare with the ground truth.

We hosted the tool in an instance of Lightsail⁸ and pulled annotators from within the same authors of the paper.

⁸<https://aws.amazon.com/lightsail/>

Q_id: 457308 | Game_id: 108574 generic_annotator_name | Logout

Target Category: skateboard | Class: category-group

Q: is it leftmost one?

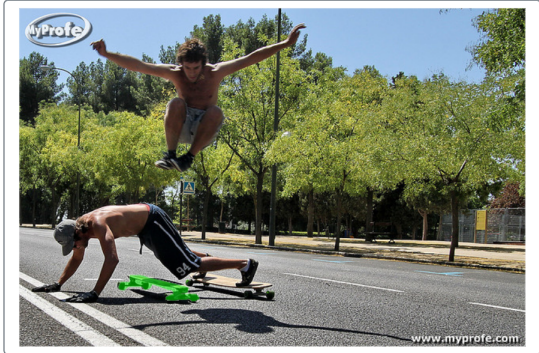
No

Maybe No

I don't know

Maybe Yes

Yes



Hide Mask

Figure 6: Screenshot of the annotation tool used.