New Datasets and Models for Contextual Reasoning in Visual Dialog (Supplementary Materials)

Yifeng Zhang, Ming Jiang, and Qi Zhao

University of Minnesota, Minneapolis MN 55455, USA
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

1 Introduction

The supplementary materials provide additional results and details of our proposed work. Specifically, they are organized as follows:

1. Sec. 2 presents details of the dataset construction method.
2. Sec. 3 presents supplementary analysis and visualization of our datasets.
3. Sec. 4 presents supplementary quantitative and qualitative results.
4. Sec. 5 presents details of the proposed NDM method.

2 Dataset Construction Details

While the main paper focuses on leveraging sampled contexts and the question engine to create diverse and context-rich visual dialogs, in this section, we present more details about the dataset construction. First, we present the details of the compounds and question templates used in the generation of dialogs. Next, we describe the details of sanity check rules applied to correct contextual errors introduced by decoy questions.

2.1 Compounds and Question Templates

In total, CLEVR-VD is constructed based on a set of 90 unique compounds and 240 question templates, while GQA-VD is constructed based on a set of 120 unique compounds and 360 question templates. Tab. 1 and Tab. 2 demonstrate the list of major compounds and sampled templates for CLEVR-VD and GQA-VD datasets. Note that for creating datasets with satisfactory diversity, each compound can be matched with one or multiple question templates.

As shown in these tables, depending on the existence of contextual dependencies, there are two types of compounds: Independent compounds such as Find-[Output], Find-Find-[Output], Find-Relate-[Output] do not depend on the dialog history. The [Output] primitive can be one of the following output modules: Count, Exist, Compare, or Describe. Therefore, questions generated from independent compounds can be answered independently. The others are more or

less dependent on other compounds because they contain the Include or Exclude primitives, such as Include-Find-[Output].

Different from existing datasets (*e.g.*, VisDial [1] and CLEVR-Dialog [3]), many of our questions have multiple contextual dependencies or depend on not only object entities but also abstract knowledge to answer. For example, Include-Find-Exclude-Count is such a complex compound, and a question generated from it can be *"How many other things share its color?"*. Such context-rich questions require VD models to establish fine-grained understanding of the dialog history, which distinguishes our datasets from previous studies. In this example, they should first query the object color, locate all objects with the same color, exclude the previously mentioned objects, and finally output the number of remaining objects.

2.2 Sanity Check Rules

To diversify the questions, we randomly replace the objects mentioned in questions with decoys. A sanity check is conducted to avoid introducing errors because of the decoys. We show a list of sanity check rules in Tab. 3. Among all the rules, Check_Existence is the first one to be executed. If a decoy does not exist in the image and the relationships/attributes still hold for the decoy in the current question, we directly proceed by replacing the pronouns of the previous object with *"the [objectname]"* across the whole dialog. If the decoy exists in the image, we first execute Check_Dependencies to check whether the context still holds for the new object, and then leverage Check_Relation and Check_Attributes to ensure the decoy-relationship and decoy-attribute integrity across the whole dialog.

3 Supplementary Data Analyses

To better illustrate different aspects of the datasets, in this section, we present quantitative dataset analyses and visualizations.

3.1 Quantitative Data Analyses

We adopt several metrics (*i.e.*, Consistency, Validity, Grounding, Plausibility) from GQA [2] to measure various aspects of our data. Although these metrics are generally designed to evaluate the performance of models, evaluation scores of the same model can be used to compare the characteristics of the datasets. We evaluate HRE-QIH [1] and MN-QIH [1], two hierarchical encoder-based or memory-based VD models on four datasets: CLEVR-Dialog, CLEVR-VD, VisDial, and GQA-VD (see Tab. 4). The evaluation scores on CLEVR-VD and GQA-VD are unanimously higher than those on CLEVR-Dialog and GQA, which suggests that our proposed datasets are of higher quality than the compared datasets in terms of question design and answer sanity.

3.2 Visualization

We visualize the CLEVR-VD and GQA-VD questions in Fig. 1 and Fig. 2, respectively. Both datasets consist of a broad range of questions. Among all terms that start questions, *"What"*, *"How"* have higher ratios because they are capable of querying different types of attributes. Compared with CLEVR-VD that has only four key attributes (*i.e.*, color, shape, material, size) and a limited number of object types, GQA-VD contains more objects/attributes. Therefore, GQA-VD lowers the ratios of those four attributes and maintains a more diverse distribution over a broader range of object categories. GQA-VD also includes more questions about the human and position. Lastly, the real-world scene in GQA enables questions about the conclusive descriptions of the whole scene.

4 Supplementary Results

In this section, we demonstrate supplementary quantitative results of baseline model performances on the VisDial ranking task [1] and more qualitative results.

4.1 VisDial Ranking Results

We compare model performances on the VisDial ranking task [1], where each model predicts the order of answer candidates. Tab. 5 shows the performances in terms of the MRR/R@k/Mean metrics. Compared with state-of-the-art VD and VQA models, NDM performs the best in all 5 metrics, demonstrating its promising performance and generalizability.

4.2 Qualitative Examples

Due to the page limit, we only present one qualitative example in the main paper. Here, Fig. 3 demonstrates four supplementary qualitative examples of our NDM method on the proposed GQA-VD and CLEVR-VD datasets. As NDM leverages our novel memory mechanism to update the memorized knowledge according to the dialog history, its reference to the entity or abstract knowledge is more accurate. On the contrary, CorefNMN directly stores all the previously attended entities in a pool and is vulnerable to wrong reference. Therefore, compared to the state-of-the-art CorefNMN method, in these examples, NDM shows better attention accuracy and reasoning performance:

In the first example, NDM shifts attention to multiple abstract concepts (*i.e.*, *"read"*, *"eat"* in Q10: *"Can anyone both eat and read at the same time?"*), while CorefNMN is unable to include such abstract knowledge.

In the second example, NDM correctly locates the referred entity (*i.e.*, *"him"* in Q7: *"What is the color of the shoes worn by him?"*), while CorefNMN finds it difficult to determine the reference among all stored entities from the dialog history.

In the third example, abstract knowledge (*i.e.*, color of *"it"* in Q7: *"Are there cylinders share the same color of it?"*) can only be transferred by NDM to locate the correct *"cylinder"*. CorefNMN locates the wrong entity *"cube"* and fails to answer correctly.

In the last example, NDM attends the correct object (*i.e.*, *"brown sphere"* as *"it"* in Q5: *"How many other objects share the same color of it?"*), while CorefNMN mistakenly refers to *"red sphere"*, producing incorrect answers for the subsequent several questions.

5 Supplementary Method

The proposed NDM belongs to the class of neural modules networks, which parse the question into a set of neural modules that characterize reasoning operations. In this section, we present detailed descriptions of the proposed NDM method, including the question parser and the objective function.

5.1 Question Parser

Generating neural modules with parameters from questions is fundamental to neural module networks. Traditionally, neural module networks of VQA tasks leverage sequential models (*e.g.*, LSTMs) to parse each question into a set of reasoning modules. Different from VQA models that reason over a single question, NDM works on context-rich dialogs, and hence should be adjusted accordingly to capture diverse contextual dependencies. For example, to parse the question *"How about its color?"* following *"What is the shape of the metallic object to the right of the image?"*, the parser should not only translate the latter question into Describe[color] module but also figure out what to Include (*i.e.*, *metal object to the right of the image*).

Therefore, as shown in Fig. 4, we design a parser that consists of two LSTMs: a question LSTM that translates questions and a memory LSTM that tracks dialog contexts. Following the XNM [4] approach, our NDM uses an LSTM-based sequence-to-sequence model to create n_j module and parameter selections $\{\boldsymbol{m}_i^j\}_{j=1}^{n_j}$ from the i-th question $\boldsymbol{q}_i$. Note that the answer $\boldsymbol{a}_i$ of the current question is also encoded into the question LSTM to update the hidden feature $\boldsymbol{h}_i'$. Apart from the question LSTM, we create a memory LSTM to track the dialog history and enable history-dependent neural module parsing. The output at step $i+1$ initializes the hidden features $\boldsymbol{h}_{i+1}$ of the corresponding question LSTM.

Table 1. A list of major compounds and corresponding sampled templates of CLEVR-VD. [AN], [Z], [C], [M], [S], [R] indicate the attribute name, size, color, material, shape, relationship of objects and the numbers indicate their index. [Z]/[C]/[M]/[S] means selecting a random attribute from all four attribute types. For the full list, please refer to `https://rb.gy/6eq0f1`.

Compound (CLEVR-VD)	Template
Find-Count	How many objects in the image? How many [Z] [C] [M] [S] objects?
Find-Filter-Count	How many objects in the image are not [Z]/[C]/[M]/[S]? How many [Z] [C] [M] [S] objects are not [Z]/[C]/[M]/[S]?
Find-Exclude-Count	How many other [Z] [C] [M] [S] are there in the picture?
Include-Find-Count	How many [Z] [C] [M] [S] among them? How many other things share its [AN]?
Include-Relate-Count	How many things to its [R]? How about to its [R]?
Find-Relate-Count	How many things are [R] that [Z] [C] [M] [S]?
Include-Find-Exclude-Count	How many other things share its [AN]? How many things have the same [AN] as that [Z] [C] [M] [S]?
Find-Exist	Are there any [Z] [C] [M] [S] in the picture? Is there any [Z] [C] [M] [S] in the picture?
Find-Filter-Exist	Are there any [Z] [C] [M] [S] in the picture that are not [Z]/[C]/[M]/[S]? Is there any [Z] [C] [M] [S] in the picture that is not [Z]/[C]/[M]/[S]?
Find-Exclude-Exist	Are there other [Z] [C] [M] [S] in the picture? Are there [Z] [C] [M] [S] among them?
Include-Relate-Exist	Are there any things to its [R]? How about to its [R]?
Find-Relate-Exist	Are there things [R] that [Z] [C] [M] [S]? Are there [Z] [C] [M] [S] objects [R] a [Z1] [C1] [M1] [S1]?
Find-Relate-Find-Exist	Are there [Z] [C] [M] [S] objects [R] a [Z1][C1][M1][S1]?
Include-Exclude-Exist	Are there other things that share its [AN]?
Include-Find-Exclude-Exist	Are there things that have the same [AN] as that [Z] [C] [M] [S]?
Include-Find-Exist	Is the [AN] [[C]/[S]/[Z]/[M]]?
Include-Find-Relate-Find-Find-Exist	Is the [AN] of previous [Z] [C] [M] [S] that is [R] to a [Z] [C] [M] [S] is also [Z]/[C]/[M]/[S]?
And(Find-Find, Find-Find) - Exist	Are [Z] [C] [M] [S] and [Z1] [C1] [M1] [S1] share the same [AN]?
Find-Describe	What is the [AN] of [Z] [C] [M] [S]?
Find-Filter-Describe	What is [AN] of [Z] [C] [M] [S] that is not [[Z1]/[C1]/[M1]/[S1]]? Name/Describe the [AN] of [Z] [C] [M] [S] that is not [Z]/[C]/[M]/[S]?
Find-Relate-Describe	What [AN] is it if there is a thing [R] that [Z] [C] [M] [S]?
Find-Relate-Find-Describe	What is the [AN] of the [Z] [C] [M] [S] object that is [R] a [Z1] [C1] [M1] [S1]?
Include-Describe	How about the [AN]? What is the [AN] of that [Z] [C] [M] [S]? What is its [AN]?
Include-Relate-Describe	What [AN] is it if there is a thing to its [R]?
Include-Include-Find-Describe	What is the [AN] of the object that shares the [AN1] of [Z1] [C1] [M1] [S1] and [AN2] of [Z2] [C2] [M2] [S2]?
And(Include-Find, Include-Find) - Find - Describe	What is the [AN] of [Z] [C] [M] [S] that shares the [A1] of [Z1] [C1] [M1] [S1] and [A2] of [Z2] [C2] [M2] [S2]?

Table 2. A list of major compounds and corresponding sampled templates of GQA-VD. Due to more diversified attribute types, we leverage [A] to denote any random attribute value. [O], [PO], [AN] indicate the object name, parent category name, attribute name, and the numbers indicate their index. Slash sign indicates choosing one from all of the candidates (*e.g.*, [O]/[PO] means selecting either a object name or its parent category name). For the full list, please refer to `https://rb.gy/6eq0f1`.

Compound (GQA-VD)	Template
Find-Count	How many [O]/[PO] are there in the image? What is the number of [O]/[PO] that are in the image?
Find-Filter-Count	How many [O]/[PO] in the image are not [A1]? What is the number of [O]/[PO] in the image that are not [A1]?
Find-Relate-Find-Count	How many [O]/[PO] are [R] to [A1] [O1] in the image? What is the number of [O]/[PO] that are [R] to [A1] [O1] in the image?
Find-Exclude-Count	How many other/rest [O]/[PO] are there in the image ? What is the number of [O]/[PO] excluding previous mentioned ones?
Include-Find-Count	How many [O]/[PO] share the [AN] of [O1] in the image? What is the number of [O]/[PO] that share the [AN] of [O1]? What is the number of [O]/[PO] that have the same [AN] as [O1]?
Include-Include-Find-Count	How many [O]/[PO] share the [A1] of [O1] and [A2] of [O2] are there in the image? What is the number of [O]/[PO] that share the [A1] of [O1] and [A2] of [O2] in the image?
Include-Find-Exclude-Count	How many other [O]/[PO] share the [AN] of it are there in the image? Apart from the mentioned [PO], what is the total number of the rest of them?
Include-Relate-Count	What is the number of [O]/[PO] that is [R] it/them?
Include-Relate-Find-Count	What is the number of [A] [O] that is [R] it/them?
Find-Exist	Are there [A] [O] in the picture? Is there any [A] [O] in the picture?
Find-Filter-Exist	Are there any [A] [O] in the picture that are not [A1]? Is there any [A] [O] in the picture that is not [A1]?
Find-Exclude-Exist	Are there other [A] [O] in the picture? Are there any [A] [O] except them?
Include-Relate-Exist	Are there things to its [R]? How about to its [R]?
Find-Relate-Exist	Are there things [R] that [A] [O]? Are there [A] [O] that is [R] a [A1] [O1]?
Find-Relate-Find-Exist	Are there [A] [O] that is [R] a [A1] [O1]?
Include-Exclude-Exist	Are there other things/objects that share its [AN]?
Include-Find-Exclude-Exist	Are there things/objects that have the same [AN] as that [A] [O]?
Include-Find-Exist	Is there any [O] shares the same [AN] of [O1]?
Include-Find-Find-Relate-Find-Find-Exist	Is there any other [A] [O] that is [R] to a [A1] [O1]
And(Find-Find, Find-Find) - Exist	Are [A] [O] and [A1] [O1] share the same [AN]?
Find-Describe	What is the [AN] of [O]?
Find-Filter-Describe	What is [AN] of [O] that is not [A]? Name/Describe the [AN] of [O] that is not [A]?
Find-Relate-Describe	What is its [AN] if there is a thing [R] that [O]?
Find-Relate-Find-Describe	What is the [AN] of the [O] object that is [R] a [O]?
Include-Describe	How about [AN]? What is the [AN] of that [A][O]? What is its [AN]?
Include-Relate-Describe	What [AN] is it if there is a thing to its [R]?
Include-Include-Find-Describe	What is the [AN] of the object that shares the [AN] of [O] and [AN1] of [O1]?
And(Include-Find, Include-Find) - Find - Describe Include-Include-Include-Describe	What is the [AN] of [O] that shares the [AN1] of [O1] and [AN2] of [O2]? Where is the scene given the [A] [O], [A1] [O1] and [A2] [O2]? What is the [AN] of the [O] given the three answered questions above?

Table 3. The list of sanity check rules. Obj/obj1/obj2 refers to the objects, D refers to the dialog, Q refers to the decoy question, rels refers to the list of relationships between objects, attrs refers to the list of attributes.

Rules	Description
Check_Existence(obj)	Check the existence of the object in the image.
Check_Dependencies(obj1, D)	Check whether the previous contextual dependencies still hold for the decoy question.
Check_Relation_Q(obj1, obj2, rels, Q)	Check whether the relation still holds for the decoy question.
Check_Attributes_Q(obj, attrs, Q)	Check whether the attributes are still compatible with the decoy questions.
Check_Relation_D(obj1, obj2, rels, D)	Check whether the relation still holds for the whole dialog.
Check_Attributes_D(obj, attrs, D)	Check whether the attributes are still compatible in the whole dialog.

Table 4. Quantitative evaluation of HRE-QIH and MN-QIH on CLEVR-Dialog, CLEVR-VD, VisDial, and GQA-VD.

Dataset	Consistency		Validity		Grounding		Plausibility	
	HRE-QIH	MN-QIH	HRE-QIH	MN-QIH	HRE-QIH	MN-QIH	HRE-QIH	MN-QIH
CLEVR-Dialog	80.99	80.72	97.13	96.95	81.28	81.33	87.36	87.53
CLEVR-VD	81.13	81.04	97.21	96.85	81.74	81.93	87.79	87.87
VisDial	71.48	72.65	94.98	94.87	79.41	80.02	83.49	83.86
GQA-VD	77.25	78.49	95.17	95.04	79.48	82.14	85.68	85.46

Table 5. Comparison between our method with SOTAs on VisDial Ranking Task. Best results are highlighted in bold.

Model	MRR	R@1	R@5	R@10	Mean
Answer Prior	0.374	23.55	48.52	53.23	26.50
NMN	0.527	40.18	69.42	74.85	9.15
BUTD	0.598	42.75	72.86	85.94	5.21
HRE-QIH	0.524	42.28	62.33	68.17	16.79
MN-QIH	0.597	45.55	76.22	85.37	5.46
CorefNMN	0.641	50.92	80.18	88.81	4.45
NDM	**0.674**	**53.62**	**81.51**	**88.37**	**4.03**

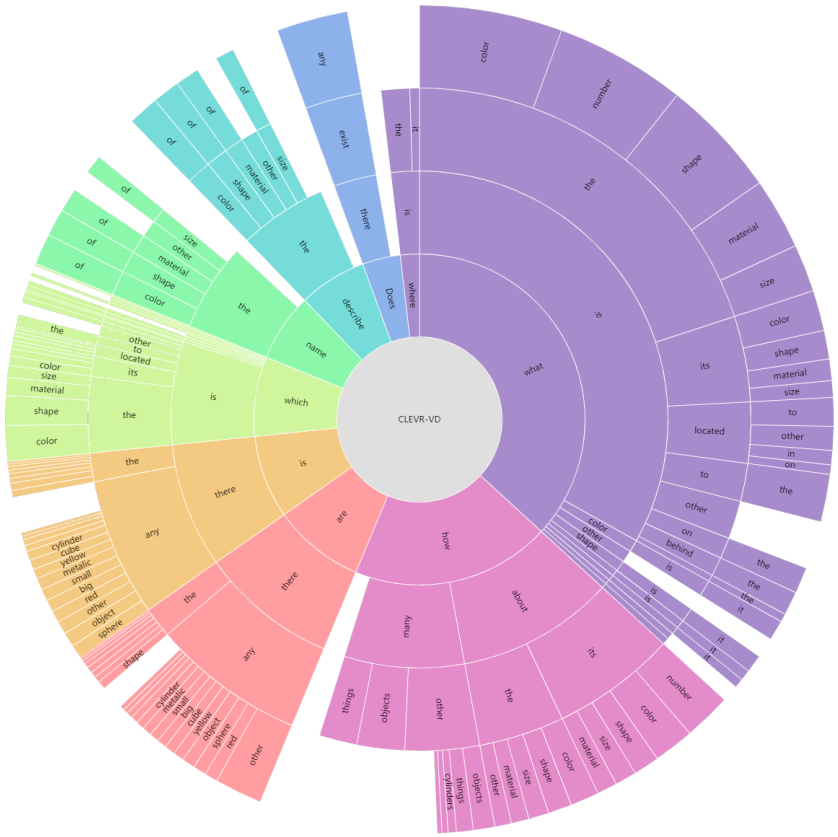

Fig. 1. Sunburst visualization for top four words of CLEVR-VD questions. Blank areas denote the omitted words that are too few to visualize.

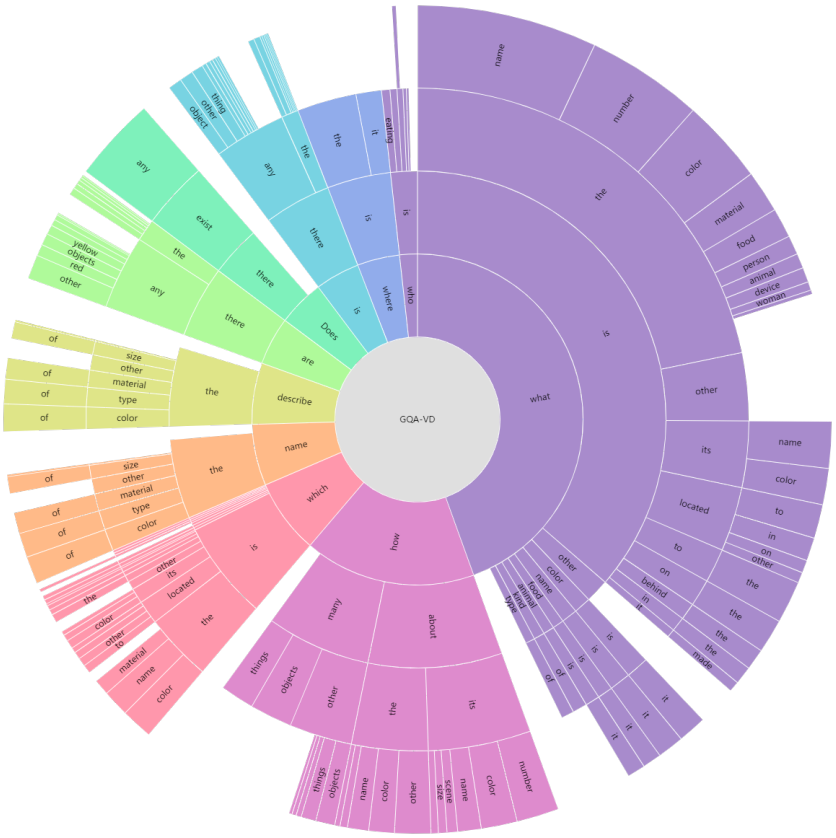

Fig. 2. Sunburst visualization for GQA-VD questions. Blank areas denote the omitted words that are too few to visualize.

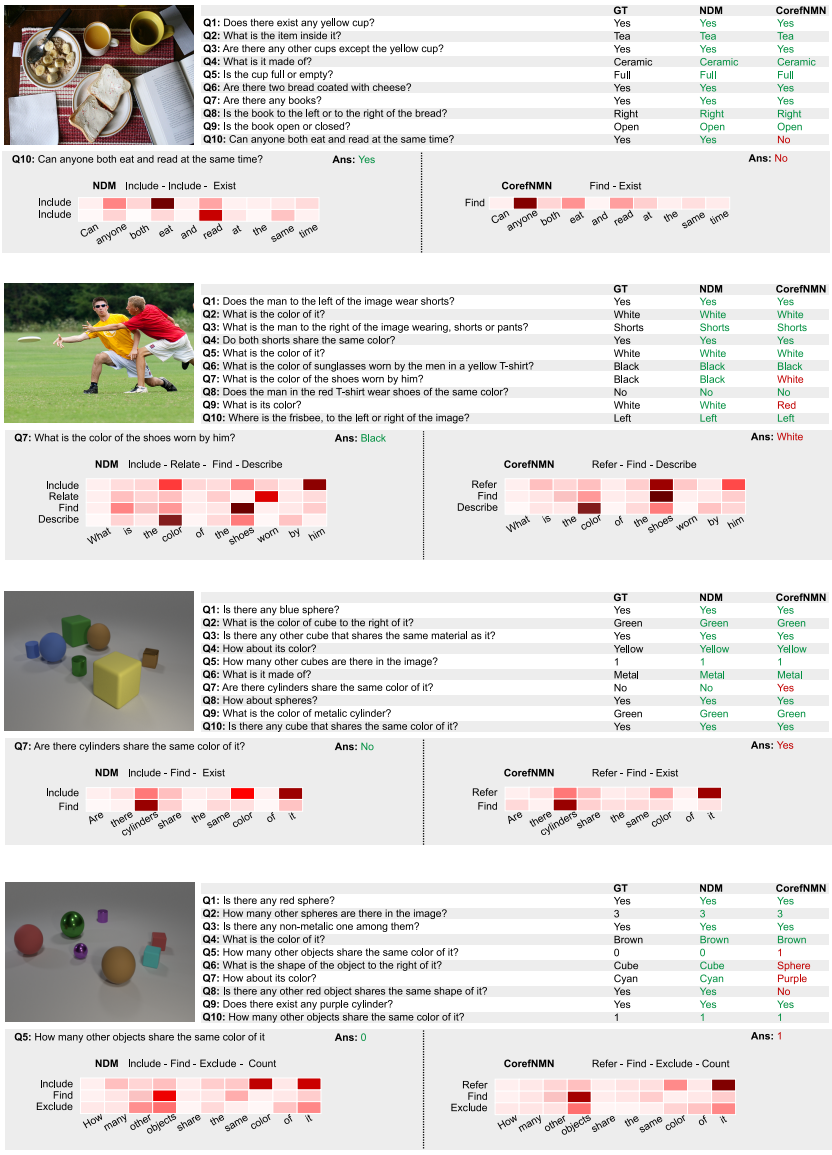

Fig. 3. Supplementary Qualitative Results. Heat maps demonstrate the attention weights at each reasoning step when answering the corresponding questions.

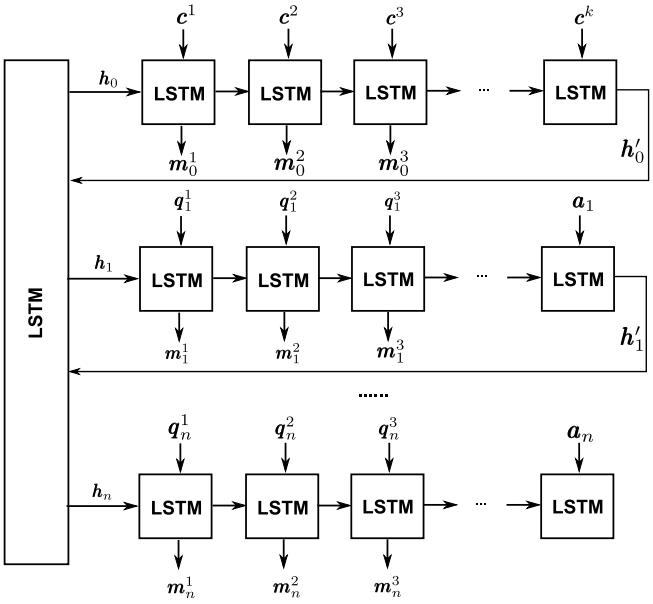

Fig. 4. Overview of the NDM's Parser. The parser consists of two nested LSTMs, where the memory LSTM (Left) encodes the dialog history and outputs the initial hidden features for the question LSTM (Right) to predict neural modules selections with the parameters.

References

1. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 326–335 (2017)
2. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6700–6709 (2019)
3. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. arXiv preprint arXiv:1903.03166 (2019)
4. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8376–8384 (2019)