New Datasets and Models for Contextual Reasoning in Visual Dialog

Yifeng Zhang, Ming Jiang, and Qi Zhao

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

Abstract. Visual Dialog (VD) is a vision-language task that requires AI systems to maintain a natural question-answering dialog about visual contents. Using the dialog history as contexts, VD models have achieved promising performance on public benchmarks. However, prior VD datasets do not provide sufficient contextually dependent questions that require knowledge from the dialog history to answer. As a result, advanced VQA models can still perform well without considering the dialog context. In this work, we focus on developing new datasets and models to highlight the role of contextual reasoning in VD. We define a hierarchy of contextual patterns to represent and organize the dialog context, enabling quantitative analyses of contextual dependencies and designs of new VD datasts and models. We then develop two new datasets, namely CLEVR-VD and GQA-VD, offering context-rich dialogs over synthetic and realistic images, respectively. Furthermore, we propose a novel neural module network method featuring contextual reasoning in VD. We demonstrate the effectiveness of our proposed datasets and method with experimental results and model comparisons across different datasets. Our code and data are available at https://github.com/SuperJohnZhang/ContextVD.

1 Introduction

Understanding vision and language and reasoning about both modalities is a challenging research problem. With the development of advanced machine learning techniques and large-scale datasets, recent progress in computer vision (CV) and natural language processing (NLP) has resulted in promising achievements in developing intelligent agents for various vision-language tasks [5,9,12,16,38]. A typical task is visual question answering (VQA) [5], which requires to answer an open-ended question about an image. As a step further, researchers generalize VQA to the more challenging visual dialog (VD) [12] task, which aims at holding a continuous question-answering dialog about visual contents. A unique challenge of VD is to understand the context of a question from the dialog history. Take the question *"what is the fruit to the right of it with the same color?"* for example (see Fig. 1) – to answer the question, one must extract contextual information from previous questions about what *"it"* and *"same color"* refer to.

To tackle this challenge, recent VD studies have developed models to keep track of all phrases in the dialog that refer to the same entity in the image (*i.e.*,

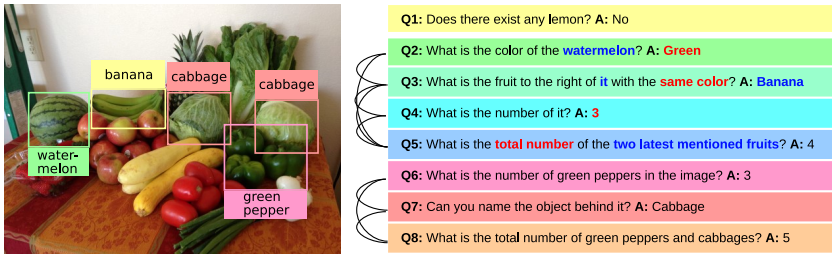

Fig. 1. An example from our GQA-VD dataset. It consists of a variety of questions that require contextual information to answer. Different from existing datasets, each GQA-VD question can refer to multiple entities (blue) or abstract concepts (red) in the dialog history, which offers a more challenging testbed for VD modeling.

coreferences) [23]. Despite their promising results in existing VD benchmarks, it has been observed that state-of-the-art VQA models can achieve comparable or better performances in some metrics (*e.g.*, mean rank) without even considering the dialog history [29]. This suggests that existing VD benchmarks place an imbalanced emphasis on answering questions that do not depend on information from the dialog context. Therefore, further advances in VD research require to bridge three research gaps in the design of VD datasets and models: 1) the unclear definition and quantification of contextual dependencies, 2) the shortage of context-dependent questions in current datasets, and 3) the lack of model design for encoding complex dialog contexts.

In this work, we bridge the research gap with new datasets and models that focus on diverse dialog contexts. Specifically, based on linguistic theories [8], we first define a hierarchy of contextual patterns that explicitly characterize **contextual dependencies**, the general and diverse relationships across different questions in a dialog. Different from visual coreferences [23] that only focus on visual entities, contextual dependencies are more general and account for a broader range of contextual relationships.

Based on the novel definitions, we then develop two context-rich VD datasets (*i.e.*, CLEVR-VD and GQA-VD) by generating dialogs based on the popular CLEVR [21] and GQA [19] datasets. Compared with existing VD datasets [12,24], our proposed datasets consist of more diverse and balanced contexts. As shown in Fig. 1, many questions of our GQA-VD dataset depend on one or multiple previous questions. They not only refer to the previously mentioned visual entities (*e.g.*, watermelon, banana, cabbage, green pepper), but also depend on the understanding of abstract concepts (*e.g.*, number, color, *etc.*). Such general and diverse contextual dependencies lead to more challenging dialogs demanding the capabilities of VD models to reason about the dialog context.

Further, we propose a neural module network approach that explicitly models the reasoning process with a novel memory design and corresponding contextual modules to enable the attention shift among the abstract contextual knowledge. Experimental results demonstrate significant improvements of our method on

the proposed datasets and existing datasets (*i.e.*, CLEVR-Dialog [24] and VisDial [12]). This work pushes the state-of-the-art VD research towards a more fine-grained and explainable direction. Our main contributions are as follows:

1. Inspired by linguistic studies, we propose a novel definition of VD based on a hierarchy of contextual patterns, explicitly characterizing how dialog contexts are involved in a dialog.

2. We propose CLEVR-VD and GQA-VD, two new VD datasets offering diverse and complex dialog contexts, enabling the development of more sophisticated VD models. We also provide structured representations of a dialog (*i.e.*, primitives, compounds, and topics) as extra annotations.

3. Based on the new definition and datasets, we propose an explainable VD method that explicitly reasons about the dialog context with a novel memory mechanism and contextual modules, resulting in significantly improved performance while demonstrating model interpretability.

2 Related Works

Language contexts in dialog. Linguistic researches have studied language contexts in dialog [6,7,8,14,36] for many decades. Linguistic theories (*e.g.*, Speech Act Theory [6,36]) have been widely applied in dialogue act classification [33,34] and dialogue state tracking [44,27]. Derived from VQA [5], the task of VD [13] performs multiround question answering. To better encode language contexts, recent VD works [20,23] consider visual coreference resolution by linking phrases and pronouns across different QA rounds that refer to the same entity. However, coreference is far from sufficient to address complex dialog contexts related to abstract concepts (*e.g.*, number or color) or multiple entities (*e.g.*, watermelon and banana in Fig. 1). Aiming to represent language contexts in a formal, mathematical, and detailed manner, we revisit the the VD task and introduce a hierarchy of dialog contextual patterns that clearly describe the semantics and functionalities of different language entities following the Speech Act Theory [6,36]. These patterns characterize a broad range of contextual dependencies.

Visual dialog datasets. VisDial [12] and CLEVR-Dialog [24] are two large-scale VD datasets for real-world and diagnostic images, respectively. To create multiround questions and answers, VisDial hires crowd workers to discuss about real-world images (*e.g.*, MSCOCO [26]), while CLEVR-Dialog leverages virtual agents to ground complete scene graphs from synthetic images (*e.g.*, CLEVR [21]). The CLEVR-Dialog has more frequent and difficult coreference cases than VisDial. We draw inspiration from CLEVR-Dialog to create our own datasets for both real-world (*e.g.*, GQA [19]) and diagnostic images (*e.g.*, CLEVR [21]). Compared to VisDial and CLEVR-Dialog, our datasets contain richer contexts in terms of both diversity and complexity. Our new datasets include a broader range of contextual dependencies other than just coreferences. The novel contextual patterns are annotated to offer detailed and structured representations that previous datasets did not provide. Another difference lies in the question generation process. Unlike CLEVR-Dialog that solely relies on

two agents to implicitly include contexts, we provide a set of randomly sampled contexts to the question engine to ensure the context diversity and complexity.

Visual dialog models. Most VD models [12,13,20,27,37,32] follow an encoder-decoder framework to fuse dialog contexts and decode either an answer ranking or free-form response. With some researches [29,10] pointing out the importance of dialog context modeling, recent works use attention networks to solve coreferences [37], and more recently, a probabilistic treatment of dialogs using conditional variational autoencoders [30] to better encode the dialog context. All those models consider coreferences implicitly by encoding features and lack interpretability. Recent studies focus on pretraining and attention modeling (*e.g.*, VisDial-BERT [31], VD-BERT [40]) to improve model performance. Different from these methods, our proposed NDM model explicitly learns the reasoning process using neural modules that result in better explainability. It is mostly relevant to the CorefNMN [23] model that learns to infer coreferences using neural module networks. Inspired by a class of explicit VQA models [4,39,41,43,18] where an instance-specific architecture is dynamically constructed from basic building blocks representing different reasoning operations, CorefNMN stores all mentioned entities in a memory and represents coreferences as a feature extraction process with novel neural module implementations. Different from CorefNMN, we develop new modules along with a memory mechanism to reason over richer contextual dependencies and achieves significant improvements.

3 Visual Dialog Context

Visual Dialog (VD) refers to the task of answering a sequence of questions about a given image in multiple rounds [13]. Understanding the context of a dialog is essential for VD models, which helps them to answer each question based on its relationship with previous ones. Although it has been well known that extracting coreferences from the dialog history can benefit the answering of new questions [37,23], existing VD models fail to demonstrate superior performance over VQA methods, because of insufficient context representation. To promote the development of context-rich VD datasets and models, in this section, we present a more structured definition of dialog contexts. Inspired by linguistic theories [6,36] and visual reasoning studies [23,25], we define dialog contexts based on three levels of basic patterns: **primitives**, **compounds**, and **topics**.

Primitives are atomic patterns derived from the Speech Act Theory [35], which also corresponds to the atomic reasoning operations defined in visual reasoning studies. For contextual reasoning in VD, we define two new primitives (*i.e.*, Include and Exclude) that represent the knowledge inclusion and exclusion through contextual dependencies. They each can refer to one or multiple concepts mentioned in previous questions, and these concepts can either be visually grounded entities or abstract ones, as specified in the parameters. Such parameters consist of 1) a list of related questions with shared knowledge, 2) the knowledge type (*e.g.*, name or number), and 3) the knowledge entity (*e.g.*, an object). In contrast, coreferences defined by previous studies (*i.e.*, visual entities

Table 1. Summary of all primitives. We introduce two novel primitives (Include, Exclude) that represent the knowledge inclusion and exclusion through contextual dependencies. [rel] – predicate in a subject-predicate-object triplet, [fea] – the feature type, (param) – the parameter of primitives (*i.e.*, a specific object or attribute), (att) – the intermediate attention map, [qids] – the IDs of related questions.

Primitive	Question	Example Compound
`Find(param)`	Where is the apple?	`Find(apple)-Describe[position]`
`Relate[rel](param)`	Which object is made of metal?	`Find(object)-Relate[madeOf](metal)`
`Filter[fea](param)`	How many objects are there excluding sphere shape?	`Find(object)-Filter[shape](sphere)-Count`
`And(att1, att2)`	What is the number of blue metal objects?	`Find(object)-And(Find(blue),Find(metal))-Count`
`Or(att1, att2)`	What is the total number of apples and bananas?	`Or(Find(apple), Find(banana))-Count`
`Not(att)`	What is the number of non-blue objects?	`Find(object)-Not(Find(blue))-Count`
`Exist`	Is there any apple?	`Find(apple)-Exist`
`Count`	What is the number of apple?	`Find(apple)-Count`
`Compare[fea](praram)`	Who is larger, the watermelon or the apple	`Find(watermelon)-Find(apple)-Compare[size](large)`
`Describe[fea]`	What is the color of apple?	`Find(apple)-Describe[color]`
`Exclude[qids][fea](param)`	How many other fruits are there in the image?	`Find(fruit)-Count-Exclude[qids][number](fruit)`
`Include[qids][fea](param)`	How many mentioned fruits are there in the image?	`Include[qids][name](fruit)-Find(prev)-Count`

that are referred to by multiple questions) can only represent a single visual entity, which is insufficient for complex contextual representation. Other primitives are defined following conventional visual reasoning operations [19], such as attention operations (*i.e.*, Find, Relate, Filter), logical operations (*i.e.*, And, Or, Not), output operations (*i.e.*, Compare, Exist, Count, Describe), *etc*. Examples of all primitives are shown in Tab. 1.

Compounds are contextual patterns composed of a sequence of primitives. Each compound corresponds to a question in the dialog. If a compound contains Include or Exclude primitives, it means that the corresponding question is dependent on previous questions in the dialog history. For instance, the question *"What is the fruit that shares the same color as the watermelon and banana?"* can be represented as a parameterized sequence of primitives Find(fruit)-Include[qids][color](watermelon, banana)- Describe[name]. Therefore, all previous questions about the watermelons, bananas or their colors are its contextual dependencies, because they share the same contextual knowledge with it.

Topics are contextual patterns defined as connected graphs of multiple questions and their dependencies. We represent questions as graph nodes and their dependencies as edges. Thus, different topics are represented as isolated graphs. Each dialog consists of at least one topic, while the maximum number of topics is the number of questions (*i.e.*, all questions are independent from each other and can be answered without knowledge from the dialog history).

The primitives, compounds, and topics defined above provide concise and informative representation of dialog contexts, which are used in Sec. 4 to ensure the contextual richness of our proposed datasets.

4 The CLEVR-VD and GQA-VD Datasets

Based on the definition in Sec. 3 and the popular visual reasoning datasets CLEVR [21] and GQA [19], we develop two novel datasets, namely CLEVR-VD

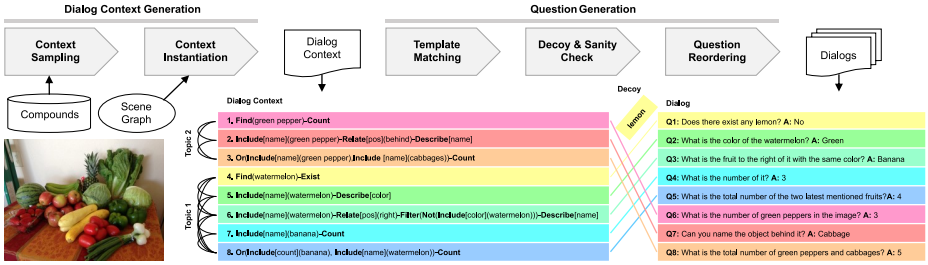

Fig. 2. An overview of the dataset generation process. First, we generate question contexts by sampling and instantiating them with parameters into a collection of topics. Next, instantiated contexts are fed into the question engine, which performs template matching, decoy and sanity check, and question reordering to generate diverse dialogs.

and GQA-VD, featuring rich dialog contexts and questions. Both datasets offer ten-round dialogs with complex contexts and diverse questions. Compared with existing VD datasets, the diversity and complexity of CLEVR-VD and GQA-VD demonstrate great potential for developing and benchmarking VD models capable of better contextual reasoning. In this section, we first describe the process of generating dialogs, and then report the data statistics.

4.1 Dataset Generation

Previous studies [12,24] develop VD datasets by either recruiting crowd workers or developing AI agents to perform question answering. Though these datasets consist of naturally generated questions about the context, there is no explicit control over the richness of contextual dependency. Differently, we generate dialogs explicitly from a structured representation of dialog context following the definition in Sec. 3. As shown in Fig. 2, the data generation process consists of five steps: context sampling, context instantiation, template matching, decoy & sanity check, and question reordering. Following these steps, we 1) generate complex dialog contexts with a variety of primitives, compounds, and topics, and 2) develop a question engine to generate a diverse set of dialogs based on each dialog context. We summarize these data generation steps in this section. For more details, please refer to the supplementary.

1. Context sampling. Different from existing datasets that generate questions directly from the scene graph of images, in this work, we aim to ensure the contextual richness of the generated dialogs. Therefore, we first randomly sample a number of predefined compounds and make sure they contain a sufficient number of contextual dependencies. These compounds specify the general layout of the dialog context without concrete parameters. In particular, for each sampled compound consisting of Include or Exclude primitives, we recursively sample their dependencies, which generates complex topics. With this approach, we arrive at a preliminary layout of the dialog context. It contains a number of

topics, each forming a graph with compounds as nodes and their dependencies as edges, indicating the overall contextual relationships of questions.

2. Context instantiation. The previous step specifies the structure of the dialog context without taking into account the visual information. Next, given the scene graph of an input image, we instantiate the dialog context by filling in the parameters. Specifically, we first randomly sample objects and attributes from the scene graph and assign them to each primitive. We then validate the compounds to makes sure that a question depending on another one must have shared parameters, but independent questions must all have different parameters. For example, when the referred object is unique in the image, a question about the object could be independent of the context. This process leads to an image-specific dialog context with rich contextual dependencies.

3. Question templates. From a dialog context, we can generate a variety of dialogs by choosing different question templates for each compound. For example, the questions *"Is there any watermelon?"*, *"Does there exist any watermelon?"* can be generated from the compound Find(watermelon)-Exist using different templates. We not only design 240 templates for CLEVR-VD and 360 templates for GQA-VD, but also prepare a set of synonyms to further increase the language diversity. The lists of templates are presented in the supplementary.

4. Decoys and sanity check. To further increase the diversity of the dialogs, we randomly replace objects or attributes in the questions with plausible decoys. The decoys do not necessarily exist in the image and they may affect the answer.After the replacement, to maintain the validity of questions, we perform sanity check based on a set of predefined rules. For example, considering the two questions *"Does there exist any watermelon?"* and *"what is the color of it?"* (see Fig. 2), with a decoy *"lemon"*, the first one may be changed to *"Does there exist any lemon?"*. Due to this change, the next question must be revised to *"What is the color of the watermelon?"* to maintain the validity of the dialog context. By making adjustments to the affected questions accordingly, these rules (see the supplementary for details) of the sanity check ensure the dialog-image integrity.

5. Question reordering. Although the order of questions has been determined by the dialog context, some questions in the dialog can be reordered without breaking the integrity of the context. For example, as shown in Fig. 2, independent questions or topics can be randomly shuffled without affecting each other, since they do not require shared knowledge. Therefore, by shuffling the question orders we further increase the diversity of dialogs.

4.2 Dataset Analysis

Tab. 2 compares the overall statistics between ours and the related VisDial [13] and CLEVR-Dialog [24]. These datasets are grouped based on their image sources: VisDial and GQA-VD use COCO images, while CLEVR-Dialog and CLEVR-VD use CLEVR images. Both CLEVR-VD and GQA-VD have several unique characteristics that distinguish them from the previous ones. For example, they have larger sizes of vocabulary and unique questions. GQA-VD has 5 times more questions than VisDial and 3 times more unique questions, making it more diverse

Table 2. Dataset statistics of CLEVR-Dialog, CLEVR-VD, VisDial and GQA-VD. Q. – questions, A. – answers, T. – topics, C. – contextual dependencies. Note that the 1.4k unique VisDial answers are short answers extracted from the 340k long answers by removing synonyms, while the 1.8k short answers of GQA-VD can also be augmented into 840k unique long answers with the current templates.

	CLEVR-Dialog	CLEVR-VD	VisDial	GQA-VD
Image Type	Synthetic	Synthetic	Real	Real
# Images	85k	100k	123k	113k
# Questions	4.25M	2M	1.2M	5.6M
# Unique Q.	73k	89k	380k	970k
# Unique A.	29	76	1.4k	1.8k
Vocab Size	125	240	7k	11k
Mean Q. Length	10.6	11.2	5.1	11.9
# T. Per Dialog	6.7	4.1	7.9	4.4
# Q. Per T.	2.3	2.9	1.4	2.6
# C. Per Q.	1.6	2.1	0.9	1.8
% Long-term C.	56	63	48	65
% Independent Q.	69	36	78	39

for mitigating biases. Although the total number of questions for CLEVR-VD is smaller than CLEVR-Dialog, it has more unique questions and answers. In particular, compared with CLEVR-Dialog and VisDial, our datasets have a reduced number of topics and more contextual dependencies per question. They also have more long-term contextual dependencies between non-adjacent questions and fewer independent questions. These statistics suggest that our datasets have more complex dialog contexts, with more questions being dependent on each other. In the following, we analyze the distribution of questions and answers, as well as different contextual patterns. Detailed statistics of our datasets are reported in the supplementary.

Balanced questions and answers. One of the main challenges of VQA and VD is the prevalent language bias [1,2,11,15,42] that allows models to answer questions based on shallow question-answer correlations rather than reasoning over both modalities. To mitigate such bias and encourage models to focus on the learning of dialog contexts, we diversify and balance the question and answer categories in the generated dialog. Fig. 3a-b show the answer distribution for the six major question categories of CLEVR-VD and the top-10 question categories of GQA-VD. As it is shown, the answers are well-balanced for each question category, which reduces the tendency of models fitting the language bias.

Diverse contextual patterns. The core characteristics of both CLEVR-VD and GQA-VD are their diverse contextual patterns. Fig. 3c demonstrates the statistics of various patterns (*i.e.*, primitives and number of contextual dependencies) for both datasets. Although their total numbers of compounds are different, CLEVR-VD and GQA-VD maintain a similar distribution of primitives and compounds. In particular, more than half of all questions have at least two contextual dependencies, which is significantly higher than existing VD datasets. The increased number of contextual dependencies leads to more challenging benchmarks for future VD models.

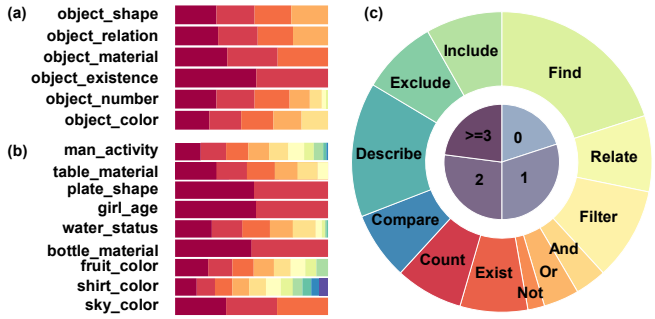

Fig. 3. Our CLEVR-VD and GQA-VD datasets maintain a balanced distribution of answers and contextual patterns. (a) Answer distribution of the six major question types of CLEVR-VD. (b) Answer distribution of the top-10 question types of GQA-VD. (c) Distribution of primitives and number of contextual dependencies.

5 Explainable Contextual Reasoning

To model the rich and diverse dialog contexts, we develop a Neural Dialog Modular network (NDM) for explainable contextual reasoning. In particular, we propose a memory mechanism and two contextual modules to explicitly store and transfer knowledge across different questions to tackle specific challenges in understanding dialog contexts. These novel components enable the shift of attention to multiple abstract concepts through diverse contextual dependencies rather than just a single coreference [23].

Neural module networks are a class of explainable reasoning methods [4,39,43]. They perform visual reasoning by first parsing the questions into a set of predefined reasoning modules to dynamically construct a network and then feeding the visual input to the network to predict an answer. Our NDM method adopts conventional question parser and VQA modules following the NMN approach [4]. Tab. 3 shows the implementation of our neural modules. In the following, we briefly present the design of our novel components: memory and contextual modules. More details are presented in supplementary.

Memorizing visual and semantic features. Due to the complexity of dialog contexts, knowledge from the dialog history can be critical for answering questions, while simply storing features of coreferences can be insufficient. For example (see Fig. 1), to answer *"What is the total number of the two latest mentioned fruits?"*, abstract knowledge (*e.g.*, the number of watermelons) can be included from the history to help answer the question. To effectively retrieve the relevant knowledge, we propose a novel memory mechanism $\boldsymbol{M}_t$ that stores both the attended visual features $\boldsymbol{M}_t^v$ and their corresponding semantic embeddings $\boldsymbol{M}_t^p$. The memory (as shown in Fig. 4) is updated by projecting the concatenation of the previous memory $\{\boldsymbol{M}_{t-1}^v, \boldsymbol{M}_{t-1}^p\}$ and current features $\{\boldsymbol{m}_t^v, \boldsymbol{m}_t^v\}$

Table 3. Implementation of neural modules. Apart from common neural modules, we design two novel contextual modules (Include, Exclude) to include or exclude the memorized features from the dialog history. MLP(·) indicates a multi-layer perceptron consisting of several fully-connected and ReLU layers, $\boldsymbol{W}_h$ is the transfer matrix computed following [43], and $\boldsymbol{W}$ is a set of K matrices of learnable weights [39] that map features onto K specific fields. $\boldsymbol{a}$, $\boldsymbol{h}$, and $\boldsymbol{q}$ indicate the input attention, features, and parameters. $\boldsymbol{a}'$ and $\boldsymbol{h}'$ are the output attention and features, respectively. $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are two input attention maps for Or/And, while $\boldsymbol{h_1}$,$\boldsymbol{h_2}$ are two input features for Compare.

Modules	Category	Operation
Or	Logic	$\boldsymbol{a}' = \max(\boldsymbol{a}_1, \boldsymbol{a}_2)$
And	Logic	$\boldsymbol{a}' = \min(\boldsymbol{a}_1, \boldsymbol{a}_2)$
Not	Logic	$\boldsymbol{a}' = 1 - \boldsymbol{a}$
Find	Attention	$\boldsymbol{a}' = \text{softmax}(\text{MLP}(\boldsymbol{h}, \boldsymbol{q}))$
Relate	Attention	$\boldsymbol{a}' = norm(\boldsymbol{W}_h \boldsymbol{a})$
Filter	Attention	$\boldsymbol{a}'$=And[$\boldsymbol{a}$, Find($\boldsymbol{q}$)]
Compare	Output	$\boldsymbol{h}' = \text{MLP}(\boldsymbol{W}(\boldsymbol{h}_1 - \boldsymbol{h}_2)))$
Count	Output	$\boldsymbol{h}' = \text{MLP}(\text{sum}(\boldsymbol{a}))$
Exist	Output	$\boldsymbol{h}' = \text{MLP}(\text{sum}(\boldsymbol{a}))$
Describe	Output	$\boldsymbol{h}' = \text{softmax}(\text{MLP}(\boldsymbol{q}))\boldsymbol{W}(\boldsymbol{a} \circ \boldsymbol{h})$
Include	Attention	Or[$\boldsymbol{a}$, softmax($Eq.$ (4))]
Exclude	Attention	And[$\boldsymbol{a}$, Not[softmax($Eq.$ (4))]]

$$\boldsymbol{M}_t^v = tanh(\boldsymbol{W}^v[\boldsymbol{M}_{t-1}^v, \boldsymbol{m}_t^v]) \tag{1}$$

$$\boldsymbol{M}_t^p = tanh(\boldsymbol{W}^p[\boldsymbol{M}_{t-1}^p, \boldsymbol{m}_t^p]), \tag{2}$$

where $\boldsymbol{W}_v$, $\boldsymbol{W}_p$ are learnable parameters. $\boldsymbol{m}_t^v$ is the duplication of current attended visual features, while $\boldsymbol{m}_t^p$ describes the attended language features by encoding the dialog history into semantic embeddings with an LSTM [17].

Contextual modules. To precisely extract relevant information from the attended visual features $\boldsymbol{M}^v$ and their semantic embeddings $\boldsymbol{M}^p$, we also implement Include and Exclude as novel contextual modules. Different from CorefNMN [23], our contextual modules extract visual features from the memory $\boldsymbol{M}^v$, project them into several feature spaces ($e.g.$, name, color, count) and finally produce the abstract features with a linear combination.

As shown in Fig. 4, given the memorized features $\boldsymbol{M}^v$, the input parameter $\boldsymbol{q}$ and the image features $\boldsymbol{h}$, we can obtain relevant features $\boldsymbol{h}_m$ from the memory

$$\boldsymbol{h}_m = \text{softmax}(\text{MLP}(\boldsymbol{M}^v, \boldsymbol{q})) \circ \boldsymbol{h}, \tag{3}$$

where $\circ$ denotes the Hadamard product. The relevant features $\boldsymbol{h}_m$ are then projected into K spaces with the same learnable projecting matrix ($\boldsymbol{W} = \{\boldsymbol{W}_k\}_{k=1}^K$) as Describe. Finally, given the memorized semantic embeddings $\boldsymbol{M}^p$ and target feature name $\boldsymbol{p}$, we measure the overlap of their probability distributions ($i.e.$,

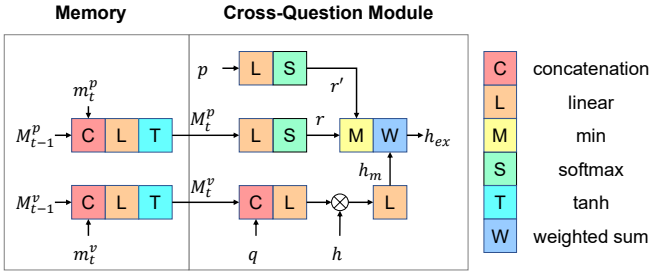

Fig. 4. The proposed memory mechanism and contextual modules that retrieve relevant knowledge $\boldsymbol{h}_{ex}$ from the dialog history. Contextual modules first find the attended features of relevant entity from image feature $\boldsymbol{h}$ (with memorized attended feature $\boldsymbol{M}_t^v$ and the parameter $\boldsymbol{q}$), and then retrieve the relevant knowledge using a weighted combination of the features projected over different spaces (*e.g.*, name, color, number). The weights are computed by measuring the overlap between the memorized semantic embedding $\boldsymbol{M}_t^p$ and target feature name $\boldsymbol{p}$.

$\boldsymbol{r} = \text{softmax}(\text{MLP}(\boldsymbol{M}_p))$, $\boldsymbol{r}' = \text{softmax}(\text{MLP}(\boldsymbol{p}))$ as weights and weighted combine K projections to obtain the extracted features

$$\boldsymbol{h}_{ex} = \sum_{k=1}^{K} \min(r_k, r'_k) \boldsymbol{W}_k \boldsymbol{h}_m, \tag{4}$$

where r_k, r'_k are the k-th entries of $\boldsymbol{r}$ and $\boldsymbol{r}'$. Finally, as shown in Tab. 3, the Include and Exclude modules process the result ($\boldsymbol{h}_{ex}$) of Eq. (4) differently to determine the inclusion or exclusion of the retrieved knowledge.

6 Experiments

Our proposed datasets provide new opportunities for developing and benchmarking context-aware VD models. In this section, we conduct extensive experiments to demonstrate the effectiveness of our datasets and the proposed NDM method. Sec. 6.2 reports quantitative results in comparison with the state-of-the-art. Sec. 6.3 visualizes the parameters of neural modules to illustrate the contextual knowledge reasoning. Sec. 6.4 analyzes the effectiveness of our novel memory mechanism and contextual modules.

6.1 Models and Evaluation

We systematically evaluate NDM and a series of baselines and state-of-the-art models. First, we develop a baseline model that predicts the answers based on the prior distribution of the training data. We then compare our method with three VD models (*i.e.*, HRE-QIH [12], MN-QIH [12], CorefNMN [23]) and two VQA models (*i.e.*, NMN [4], BUTD [3]). In addition, we incorporate pretrained

Table 4. Quantitative comparison with state-of-the-art methods on CLEVR-Dialog, CLEVR-VD, VisDial, and GQA-VD datasets.

Model	CLEVR-Dialog	CLEVR-VD	VisDial	GQA-VD
Answer Prior	33.42	27.52	23.55	30.06
NMN [4]	56.63	45.47	42.18	52.18
BUTD [3]	65.74	50.85	46.75	52.90
HRE-QIH [12]	63.38	57.41	42.28	55.97
MN-QIH [12]	59.65	54.96	45.55	57.75
CorefNMN [23]	68.03	56.82	50.92	56.59
NDM	**68.21**	**59.89**	**52.72**	**60.84**
VD-BERT [40]	68.12	59.67	51.63	60.67
VisDial-BERT [31]	68.20	59.78	**53.85**	60.89
NDM-BERT	**68.23**	**59.92**	52.91	**61.08**

ViLBERT [28] features into our NDM model, and compare it (*i.e.*, NDM-BERT) with language-pretrained VD-BERT [40] and VisDial-BERT [31] methods. We train and evaluate these models on our proposed CLEVR-VD and GQA-VD datasets, as well as two public datasets: CLEVR-Dialog [24] and VisDial [12]. All the compared models are trained with default parameters, and evaluated on the validation sets. Our NDM and NDM-BERT models are optimized using the Adam [22] optimizer with a learning rate of 10^{-4} and a decay rate of 10^{-5}.

6.2 Quantitative Results

Tab. 4 shows quantitative results demonstrating the importance of context-rich datasets for visual dialog modeling. In general, we find that the VD models perform much better on CLEVR-VD and GQA-VD than VQA models (*i.e.*, NMN and BUTD in the top panel), suggesting that the more challenging dialogs of our datasets with complex contextual patterns cannot be handled without reasoning about contextual dependencies. Further, we find that our NDM achieves the highest accuracy among all non-pretraining methods (*i.e.*, HRE-QIH, MN-QIH, CorefNMN). The significant gains on CLEVER-VD and GQA-VD datasets demonstrate its ability to reason about rich dialog contexts, and its high performances on CLEVR-Dialog and VisDial demonstrate our model's generalizability.

Though NDM is a neural module network focusing on structured reasoning but not pretraining, Tab. 4 also compares it with the state-of-the-art methods based on language pretraining (the bottom panel of Tab. 4). The proposed NDM, without pretraining, is competitive among the state-of-the-art pretrained models. It also consistently outperforms VD-BERT on all four datasets. Further, our pretrained NDM-BERT maintains interpretability while achieving the best performance (*i.e.*, also outperforming VisDial-BERT) on CLEVR-Dialog, CLEVR-VD, and GQA-VD. Between NDM and NDM-BERT, we only observe minor performance improvements, which suggests that the learning of contextual dependencies does not benefit significantly from pretraining.

Tab. 5 groups the questions into categories with different numbers of contextual dependencies and shows the average accuracy for each category. It is

Table 5. Average accuracy for questions with different numbers of contextual dependencies on CLEVR-VD and GQA-VD.

	CLEVR-VD				GQA-VD			
Model	0	1	2	≥3	0	1	2	≥3
Answer Prior	28.75	28.24	26.53	26.65	31.99	31.71	28.23	28.34
NMN [4]	48.36	45.79	44.28	43.94	57.86	52.33	49.92	49.68
BUTD [3]	60.95	48.90	48.12	47.82	61.53	51.79	50.38	49.78
HRE-QIH [12]	**61.95**	59.85	54.67	53.49	60.62	56.88	53.74	53.30
MN-QIH [12]	60.48	55.16	52.74	52.53	**61.85**	58.52	55.60	54.69
CorefNMN [23]	60.83	58.78	54.31	53.74	60.76	59.45	53.79	52.51
NDM	60.52	**60.13**	**59.66**	**59.32**	61.28	**61.47**	**60.61**	**59.92**

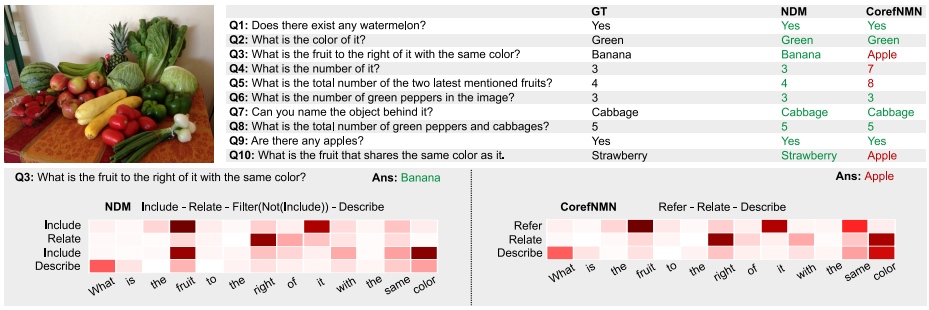

Fig. 5. A typical example on the GQA-VD dataset. Heat maps demonstrate the attention of each parameterized reasoning module when answering Q3.

noteworthy that for VQA models, the performances decrease significantly with the number of contextual dependencies, while for VD models the performance drop is less significant. Our proposed NDM performs almost equally well on questions with different number of dependencies, suggesting its ability to perform contextual reasoning across multiple questions.

6.3 Qualitative Analysis

Fig. 5 shows a typical example of answering questions in a context-rich dialog, with attention maps demonstrating the reasoning processes of NDM and CorefNMN. In this dialog, NDM shifts attention to multiple abstract concepts in the contextual knowledge, while CorefNMN only focuses on visual entities. The dialog starts with questions about the existence and color of the watermelon, and both models answer correctly. However, CorefNMN fails to answer Q3 and the subsequent questions Q4 and Q5 that depend on Q3. It incorrectly answers *"apple"* that is also to the right of the watermelon, but with different colors. Differently, NDM correctly locates the banana that is both *"to the right of the watermelon"* and *"with the same color"*. It is because our NDM can acquire both the name and color of the watermelon. By memorizing this knowledge and

Table 6. Ablation study of CorefNMN [23] and NDM baselines with different combinations of conventional VQA modules, memory (M), and contextual modules (C).

Model	CLEVR-VD	GQA-VD
CorefNMN (VQA)	54.69	54.06
CorefNMN (VQA + M)	56.45	56.24
CorefNMN (VQA + C)	56.71	56.46
CorefNMN (VQA + M + C)	**57.72**	**58.87**
NDM (VQA)	55.27	54.95
NDM (VQA + M)	57.81	57.63
NDM (VQA + C)	58.12	57.98
NDM (VQA + M + C)	**59.80**	**60.84**

leveraging multiple contextual dependencies, NDM performs more effectively in reasoning across questions. The ability of using multiple Include modules to infer complex contextual dependencies allows NDM to focus on the watermelon and its color in different reasoning steps, while CorefNMN fails to handle such complexity. Further qualitative results are reported in the supplementary.

6.4 Ablation Study

To analyze the contributions of different technical components, we further compare NDM variants with different combinations of conventional VQA modules, contextual modules (C) and the memory mechanism (M). Similarly, we adapt CorefNMN by keeping its original VQA neural modules but replacing its coreference modules and/or its memory mechanism with ours. Tab. 6 shows the results on the CLEVR-VD and GQA-VD datasets. We find that our memory and contextual modules contribute significantly to the model accuracy, leading to further improvements when they are combined. They are shown to be general, with consistent performance gains on both baselines.

7 Conclusion

Research on VD could fundamentally change the experience of human-machine interaction. However, VD studies are limited by insufficient contextual dependencies in existing datasets. To overcome this limitation, we introduce a novel definition of the dialog context with a hierarchy of contextual patterns, and construct two new VD datasets, CLEVR-VD and GQA-VD. We further propose NDM, a neural module network that performs explainable visual reasoning over the dialog context across different questions. Experimental results demonstrate that our proposed datasets offer a more general and challenging benchmark for VD models. Our NDM method also achieves promising performance by explicitly memorizing and retrieving contextual knowledge. We hope that our work will inspire future developments of interpretable and contextual reasoning methods.
Acknowledgment: This work is supported by NSF Grants 1908711 and 1849107.

References

1. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1955–1960 (2016)
2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4971–4980 (2018)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
4. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 39–48 (2016)
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
6. Austin, J.L.: How to do things with words. Oxford university press (1975)
7. Brooks, N.: Language and language learning, theory and practice (1964)
8. Bühler, K.: Theory of language. The representational function of language (1990)
9. Chattopadhyay, P., Yadav, D., Prabhu, V., Chandrasekaran, A., Das, A., Lee, S., Batra, D., Parikh, D.: Evaluating visual conversational agents via cooperative human-ai games. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 5 (2017)
10. Chen, F., Chen, X., Meng, F., Li, P., Zhou, J.: Gog: Relation-aware graph-over-graph network for visual dialog. arXiv preprint arXiv:2109.08475 (2021)
11. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017)
12. 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)
13. Das, A., Kottur, S., Moura, J.M., Lee, S., Batra, D.: Learning cooperative visual dialog agents with deep reinforcement learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2951–2960 (2017)
14. Giles, H., Coupland, N.: Language: Contexts and consequences. Thomson Brooks/Cole Publishing Co (1991)
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
16. Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., Feris, R.: Dialog-based interactive image retrieval. Advances in neural information processing systems **31** (2018)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
18. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. Advances in Neural Information Processing Systems **32** (2019)

19. 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)
20. Jiang, X., Yu, J., Qin, Z., Zhuang, Y., Zhang, X., Hu, Y., Wu, Q.: Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11125–11132 (2020)
21. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: Visual coreference resolution in visual dialog using neural module networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 153–169 (2018)
24. 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)
25. Li, M., Moens, M.F.: Modeling coreference relations in visual dialog. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 3306–3318. Association for Computational Linguistics, Online (Apr 2021), https://www.aclweb.org/anthology/2021.eacl-main.290
26. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
27. Liu, B., Tür, G., Hakkani-Tür, D., Shah, P., Heck, L.: Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In: Proceedings of NAACL-HLT. pp. 2060–2069 (2018)
28. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
29. Massiceti, D., Dokania, P.K., Siddharth, N., Torr, P.H.: Visual dialogue without vision or dialogue. arXiv preprint arXiv:1812.06417 (2018)
30. Massiceti, D., Siddharth, N., Dokania, P.K., Torr, P.H.: Flipdial: A generative model for two-way visual dialogue. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6097–6105 (2018)
31. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
32. Murahari, V., Chattopadhyay, P., Batra, D., Parikh, D., Das, A.: Improving generative visual dialog by answering diverse questions. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019)
33. Raheja, V., Tetreault, J.: Dialogue Act Classification with Context-Aware Self-Attention. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3727–3733 (2019)
34. Ravi, S., Kozareva, Z.: Self-governing neural networks for on-device short text classification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 887–893 (2018)

35. Searle, J.R., Kiefer, F., Bierwisch, M., et al.: Speech act theory and pragmatics, vol. 10. Springer (1980)
36. Searle, J.R., Searle, J.R.: Speech acts: An essay in the philosophy of language, vol. 626. Cambridge university press (1969)
37. Seo, P.H., Lehrmann, A., Han, B., Sigal, L.: Visual reference resolution using attention memory for visual dialog. Advances in neural information processing systems **30** (2017)
38. Shekhar, R., Baumgärtner, T., Venkatesh, A., Bruni, E., Bernardi, R., Fernandez, R.: Ask no more: Deciding when to guess in referential visual dialogue. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1218–1233 (2018)
39. 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)
40. Wang, Y., Joty, S., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.: VD-BERT: A unified vision and dialog transformer with bert. arXiv preprint arXiv:2004.13278 (2020)
41. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
42. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing and answering binary visual questions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5014–5022 (2016)
43. Zhang, Y., Jiang, M., Zhao, Q.: Explicit knowledge incorporation for visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1356–1365 (June 2021)
44. Zhong, V., Xiong, C., Socher, R.: Global-locally self-attentive encoder for dialogue state tracking. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1458–1467 (2018)