Mind the Context: The Impact of Contextualization in Neural Module Networks for Grounding Visual Referring Expressions

Arjun R. Akula[1], **Spandana Gella**[2], **Keze Wang**[1], **Song-Chun Zhu**[4,5,6], **Siva Reddy**[3]
[1]UCLA Center for Vision, Cognition, Learning, and Autonomy, [2]Amazon Alexa AI
[3]Facebook CIFAR AI Chair, Mila; McGill University
[4]Beijing Institute for General Artificial Intelligence (BIGAI),
[5]Tsinghua University, [6]Peking University
aakula@ucla.edu, sgella@amazon.com, kezewang@gmail.com
s.c.zhu@pku.edu.cn, siva.reddy@mila.quebec

Abstract

Neural module networks (NMN) are a popular approach for grounding visual referring expressions. Prior implementations of NMN use pre-defined and fixed textual inputs in their module instantiation. This necessitates a large number of modules as they lack the ability to share weights and exploit associations between similar textual contexts (e.g. "*dark cube on the left*" vs. "*black cube on the left*"). In this work, we address these limitations and evaluate the impact of contextual clues in improving the performance of NMN models. First, we address the problem of fixed textual inputs by parameterizing the module arguments. This substantially reduce the number of modules in NMN by up to 75% without any loss in performance. Next we propose a method to contextualize our parameterized model to enhance the module's capacity in exploiting the visiolinguistic associations. Our model outperforms the state-of-the-art NMN model on CLEVR-Ref+ dataset with +8.1% improvement in accuracy on the single-referent test set and +4.3% on the full test set. Additionally, we demonstrate that contextualization provides +11.2% and +1.7% improvements in accuracy over prior NMN models on CLOSURE and NLVR2. We further evaluate the impact of our contextualization by constructing a contrast set for CLEVR-Ref+, which we call CC-Ref+. We significantly outperform the baselines by as much as +10.4% absolute accuracy on CC-Ref+, illustrating the generalization skills of our approach. Our dataset is publicly available at https://github.com/McGill-NLP/contextual-nmn.

1 Introduction

Visual referring expression recognition is the task of identifying the object in an image that is referred to by a natural language expression (Kazemzadeh et al., 2014; Mao et al., 2016). It is a fundamental language-to-vision matching problem and has several downstream applications such as question answering, robot navigation, and image retrieval (Zhu et al., 2016; Qi et al., 2020; Young et al., 2014; Yang et al., 2016; Tu et al., 2014; Qi et al., 2015; Liu et al., 2016; Akula and Zhu, 2019; Akula, 2015; Palakurthi et al., 2015). Recently, neural module networks (NMN; Andreas et al. 2016b; Hu et al. 2017b; Liu et al. 2019) have been gaining popularity as a promising approach for solving this task. Briefly, NMN models use an explicit modular reasoning process where a program generator first analyzes the input referring expression and predicts a sequence of learnable *neural modules* (e.g. `count`, `filter`, `compare`). Next, an execution engine dynamically assembles these modules to predict the target object in the image. Such a module based hierarchical reasoning process helps NMNs in providing high model interpretability and therefore facilitates in improving overall trust in the model (Andreas et al., 2016b; Akula et al., 2020b).

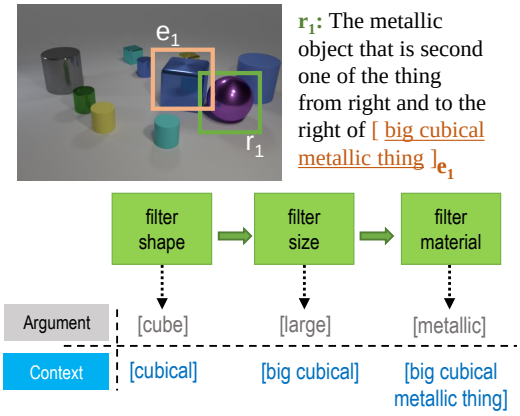


Figure 1: An example from the CLEVR-Ref+ dataset. In addition to passing textual inputs (arguments) `cubical`, `large` and `metallic` to neural modules, we also provide them with the relevant neighborhood of arguments as context (highlighted in blue).

Although achieving promising results, exist-

ing NMN models primarily focused on designing module architectures with textual inputs directly hard-coded in the module instantiation (Johnson et al., 2017b; Liu et al., 2019). For example, processing the textual inputs '*red*' and '*blue*' require the instantiation of two different modules `filter_color[red]` and `filter_color[blue]`. However, such a design demands a large number of learnable modules (and network parameters) and they cannot share weights for similar contextual textual inputs (e.g. '*dark cube*' vs. '*black cube*', '*shiny cylinder*' vs. '*metallic cylinder*'). Lack of these contextual signals leads to poor generalization performance on unseen but known language contexts (Lake and Baroni, 2018; Bahdanau et al., 2019).

Moreover, in the prior implementations of NMN such as IEP-Ref (Johnson et al., 2017b; Liu et al., 2019), the modules in execution engine are not conditioned on the surrounding context of their textual input in the expression. This is problematic as the modules are not given the opportunity to watch the neighborhood of textual input that helps in extracting the informative visiolinguistic context from the module's visual input. For example, the module `filter_color[dark]` needs to pick a black colored cube or a red-colored cube depending on the neighborhood context in the expression (e.g. "*the dark thing that is hardly visible*" vs. "*the dark thing among the red cubes*") and the type of cubes available in its visual input. Few implementations of NMN such as FiLM (Perez et al., 2018) and N2NMN (Hu et al., 2017a) parametrize the surrounding context of their textual input. However, the visiolinguistic context in these modules is rather shallow as they cannot jointly co-attend over potential objects of interest directly from the visual input and textual inputs.

In this work, we address the aforementioned issues and evaluate the impact of contextual signals in improving the performance of NMN models. First, we address the problem of hard-coded language inputs by parameterizing the module arguments (Figure 1), i.e., for example, we treat "`filter_size`" module as parameterized by textual input "*large*" instead of as a standalone function "`filter_size[large]`" (§3). We show that module parametrization reduces the total number of learnable modules by 75% without affecting the performance of NMNs.

Second, we use the ground-truth annotations in CLEVR-Ref+ (Liu et al., 2019), a challenging synthetic referring expression dataset, to show the evidence that providing the relevant neighborhood context of the textual input to the neural module (see Figure 1) is beneficial for improving the model's grounding performance (§4.1). We next propose a contextualization method to learn to select the most relevant neighborhood context by jointly co-attending on visual and textual inputs, eliminating the need for ground-truth contextual information (§4.2).

Our experimental results show that our approach is effective in capturing visiolinguistic relations and contextual dependencies, especially when the textual inputs are long, and has complex linguistic structures. We demonstrate that our proposed method significantly improves the performance of NMN (§5.4) in grounding visual referring expressions. Specifically, on CLEVR-Ref+ benchmark, we outperform competing NMN approaches such as IEP-Ref, FiLM and N2NMN by as much as +8.1% accuracy on single-referent split (S-Ref) and +4.3% on full-referent split (F-Ref). Additionally, we also test our approach on CLOSURE (Bahdanau et al., 2019) and NLVR2 (Suhr et al., 2019) benchmarks. CLOSURE is a VQA benchmark consisting of CLEVR-like questions with emphasis on simple and complex referring expressions. NLVR2 is a language grounding task where the goal is to determine whether an expression is true based on two paired real images. Our approach significantly outperforms the existing NMN approaches with +11.2% and +1.7% improvements in accuracy on CLOSURE and NLVR2 respectively.

We further evaluate the impact of our contextualization by constructing a set of contrasting perturbations around CLEVR-Ref+ test instances (Gardner et al., 2020), and call our new dataset CC-Ref+ (§5.6). We significantly outperform the state-of-the-art models by as much as +10.4% absolute accuracy on CC-Ref+.

2 Related Work

Referring Expression Recognition. Visual referring expression recognition (REF) is the task of identifying the object in an image that is referred to by a natural language expression (Mao et al., 2016; Kazemzadeh et al., 2014). Datasets containing real images and expressions such as RefCOCO+ (Kazemzadeh et al., 2014) and RefCOCOg (Mao et al., 2016) have been proposed to

evaluate the progress on this task. Multi-modal transformers (Lu et al., 2019; Li et al., 2019; Tan and Bansal, 2019), using pretrain-then-transfer approach, have shown superior performance on these datasets. However, these models fail to learn robust visio-linguistic contextual representations and are shown to exploit the imbalanced distribution in the train and test splits (Akula et al., 2020a; Cirik et al., 2018). Recently, CLEVR-Ref+ (Liu et al., 2019) has been introduced as a synthetic diagnostic benchmark that allows control over dataset bias. There are nearly 0.8M referring expressions of which 32% of expressions refer to only a single object (Single-referent) and 68% refer to more than one object (Multi-referent). In this paper, we refer to the full dataset as F-Ref and the single-referent subset as S-Ref. Module network (Liu et al., 2019; Johnson et al., 2017a; Andreas et al., 2016b) based architectures achieved new state-of-the-art performance on this dataset.

Neural Module Networks. Neural module networks (NMNs) learn to parse textual expressions as executable programs composed of learnable *neural modules* (Andreas et al., 2016b; Johnson et al., 2017a,b; Hu et al., 2017a). Each of these modules are specialized to compute basic reasoning tasks and can be assembled to perform complex and compositional reasoning. (Andreas et al., 2016b) used dependency trees (Zhu et al., 2013) to generate the execution layouts. (Andreas et al., 2016a) proposed dynamic NMNs that learns and adapts the structure of the execution layouts to the question. (Johnson et al., 2017b) proposed homogeneous (IEP) and generic neural modules, unlike fixed and hand-crafted neural module, in which the semantics of each neural module is learnt during training. IEP model achieves promising performance on CLEVR dataset. (Liu et al., 2019) proposed IEP-Ref by extending IEP model to CLEVR-Ref+ dataset and outperformed all the prior works. Although, compositional by design, the visiolinguistic context in these modules is rather shallow and fail to ground novel combinations of known linguistic constructs (Bahdanau et al., 2019). The major difference between our work and these prior works of NMN is that we explicitly parametrize and contextualize the neural modules by jointly attending over the visual and textual inputs.

	Modules
Unary	*Filter Shape*, *Filter Color*, *Filter Material*, *Filter Visible*, *Filter Size*, *Filter Ordinal*, *Unique*, *Relate*, *Same Size*, *Same Shape*, *Same Color*, *Same Material*, *Scene*
Binary	*Intersect*, *Union*

Table 1: Modules in Parameterized IEP-Ref

3 Module Parameterization in NMN

We propose parametrization as the first step to enable weight sharing and exploiting associations between similar textual contexts. Specifically, we evaluate the effectiveness of parameterizing module textual inputs using IEP-Ref (Liu et al., 2019) as the baseline NMN implementation. IEP-Ref, a NMN solution based on IEP (Johnson et al., 2017b), is the current state-of-the-art model on CLEVR-Ref+ dataset.[1] As shown Figure 2(a), the neural modules in IEP-Ref are represented using a standard Residual Convolution Block (RCB). Formally, each RCB module (f_n) of arity n receives n feature maps ($\mathbf{F_i}$) of shape $128 \times 20 \times 20$ and outputs a same-sized tensor $f_o = f_n(\mathbf{F_1}, \mathbf{F_2}, ..., \mathbf{F_n})$.

We parameterize each RCB module m as follows: (a) we feed all the words in the textual input e_m into an LSTM; (b) The last hidden state of LSTM h_t is then used to perform element-wise multiplication with the output of the first convolution layer in the RCB block to produce joint representation c_m of module's textual input (e_m) and visual input (v_m), which is then passed to ReLU function (see Appendix Figure 1b):

$$\begin{aligned} \mathbf{h_t} &= \text{LSTM}\left(e_{m,t}, \mathbf{h_{t-1}}\right), \\ \mathbf{c_m} &= \text{conv}\left(v_m\right) \odot \mathbf{h_t}. \end{aligned} \tag{1}$$

Table 2 shows the count of distinct modules and the model performance before and after parameterizing the RCB modules (i.e. IEP-Ref vs P-Ref). As we can see, there are total 60 distinct modules in IEP-Ref. After parameterization, the distinct number of modules reduce by 75% (i.e., 15 distinct modules) without any drop in the model performance. Table 1 presents the list of all the 15 modules in our parameterized NMN model.

In addition to evaluating the model performance on the full CLEVR-Ref+ dev (**F-Dev**) and test (**F-Test**) splits, we also evaluate the model on single-referent (S-Ref) dev (**S-Dev**) and test (**S-**

[1] We used the IEP-Ref implementation provided at the link https://github.com/ruotianluo/iep-ref

	#modules	#param. per module	F-Dev	F-Test	S-Dev	S-Test
IEP-Ref (18K programs)	60	442,752	80.54	78.20	49.89	51.50
P-Ref	15	574,336	**81.23**	**78.31**	**51.60**	**51.57**

Table 2: Count of modules, parameters and performance of IEP-Ref and parameterized model (P-Ref).

Test) splits.[2] Moreover, although the network parameters of each parameterized module slightly increase due to the additional LSTM unit, since each module in IEP-Ref can have multiple instantiations for the same textual input, we have fewer parameters than IEF-Ref in total (see Sec 5.4.1 for more discussion).

4 Contextualization in NMN

4.1 Using Ground-Truth Annotations

We extend our parameterized model by contextualizing it with the neighborhood context of textual input in the referring expression. Figure 1 shows an example. We leverage the ground-truth annotations available in CLEVR-Ref+ to provide neighborhood context for the modules as follows: Let us denote the ground-truth neural modules as m_1, m_2, m_3, ..., m_n for a given input referring expression q. Suppose the modules m_j and m_k are children for the parent module m_i in the ground-truth execution tree. We modify the architecture of each neural module shown where we concatenate the ground-truth arguments of all the children modules m_j and m_k and pass it as the neighborhood context to the parent module m_i (see Appendix Figure 1c). We test if this contextualization helps.

As an ablation, we also test the model performance where the entire expression q is provided as neighborhood context for the modules instead of the relevant neighborhood. Table 3 shows the results. Using the entire expression as the neighborhood context did not show any improvements in the model performance, perhaps due to the difficulty in searching and extracting relevant context from long CLEVR-like expressions. On the other hand, providing ground-truth neighborhood context shows significant improvement in the performance (1.71% on F-test and 3.19% on S-Test), indicating that model is able to extract informative visiolinguistic clues. Since the ground-truth human annotations are costly and difficult to obtain, we next propose a

[2]For results in the last two columns of Table 2, we trained our model using S-Ref train split.

Model	F-Dev	F-Test	S-Dev	S-Test
P-Ref	81.23	78.31	51.60	51.57
P-Ref + Input Expr.	81.10	77.01	50.88	51.45
P-Ref + GT Neighb.	**82.60**	**80.02**	**55.22**	**54.76**

Table 3: Performance of contextualized NMN models.

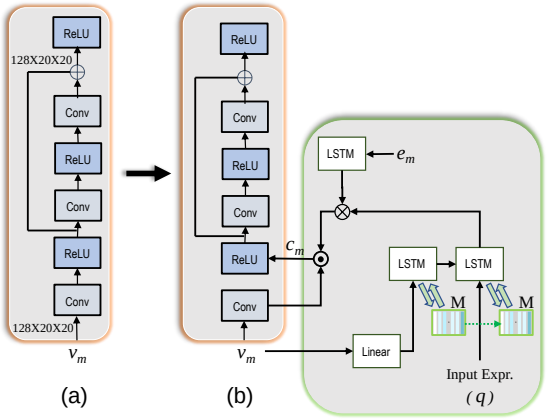


Figure 2: (a) Architecture of neural module (m) in IEP-Ref consuming a visual input v_m. $\oplus$ denotes summation. (b) Our proposed contextualized module design using our proposed memory (M) based architecture. $\odot$ and $\otimes$ denote element-wise multiplication and concatenation respectively. e_m is parameterized textual input.

contextualization method that enables the modules to learn to select the most relevant neighborhood context without requiring ground-truth annotations.

4.2 Using Memory-augmented Block

We incorporate a memory-augmented LSTM block (Graves et al., 2014) in the neural module to guide the attention towards the relevant and informative neighborhood words in the input expression (q). Figure 2(b) shows our contextualized module architecture. Our design enhances the module's capacity to exploit the visiolinguistic context between the visual input v_m and the selective set of words that are stored in the memory over multiple timesteps.

The memory M consists of a set of row vectors as memory slots. LSTM (i.e., controller) has read and write heads into M, which helps in retrieving representations from M or place them into M. In the first time step (t_0), we feed visual input and then in the later time steps textual input is fed. More formally, given a input referring expression q, at each time step (t), LSTM produces a key, $k_{i,t}$, which is either used to retrieve a particular location l from the row M_t or to store in M_t. We feed the referring

expression q into LSTM as:

$$\mathbf{h_t} = \text{LSTM}\left(\mathbf{q_t}, \mathbf{h_{t-1}}\right). \quad (2)$$

We then compute the cosine similarity measure between h_t and each individual row j in M:

$$K\left(\mathbf{h_t}, \mathbf{M_t}\left(j\right)\right) = \frac{\mathbf{h_t} \cdot \mathbf{M_t}\left(j\right)}{\|\mathbf{h_t}\| \left\|\mathbf{M_t}\left(j\right)\right\|}. \quad (3)$$

A read weight vector w_t is computed using a softmax over the cosine similarity and then a memory row m_t is retrieved. The vectors m_t, h_t are concatenated with the textual input (e_m) and then an element-wise multiplication is performed with the output of the convolution layer before passing to the ReLU function (see Appendix A).

5 Experiments

5.1 Datasets

We evaluate our approach on F-Ref and S-Ref splits of CLEVR-Ref+ (Liu et al., 2019). In addition, we also test our approach on CLOSURE (Bahdanau et al., 2019) and NLVR2 (Suhr et al., 2019) benchmarks. CLOSURE is a VQA benchmark, consisting of synthetically generated image and question pairs with emphasis on grounding simple and complex referring expressions. NLVR2 is a language grounding task where the goal is to determine whether an expression is true based on two paired real images. While reporting results on CLOSURE, we train our NMN model using CLEVR (Johnson et al., 2017a) train and val splits.

5.2 Baselines

We compare the performance of our approach against the following baselines: (1) **IEP-Ref** (Liu et al., 2019) is the current state-of-the-art NMN model for CLEVR-Ref+ benchmark which uses explicit program generator and execution engine (PG+EE) to predict the answer; (2) **FiLM** (Feature-wise Linear Modulation) (Perez et al., 2018) is a NMN model which introduces new layers in the RCB block that learn parameters $\gamma_{i,c}$ and $\beta_{i,c}$ for scaling up or down the CNN activations ($F_{i,c}$) by conditioning on the input referring expression x_i, i.e. $FiLM(\mathbf{F_{i,c}}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\mathbf{F_{i,c}} + \beta_{i,c}$; (3) **MAC** (Hudson and Manning, 2019) is an end-to-end differentiable architecture designed to perform an explicit multi-step reasoning process by decomposing them into a series of attention-based reasoning steps; (4) **VectorNMN** (Bahdanau et al., 2019) is a direct extension to FiLM that uses vector-valued inputs and outputs for the modules instead of high-capacity 3D tensors; (5) **NS-VQA** (Yi et al., 2018) uses structural scene representation from input image in addition PG+EE components in IEP-Ref; (7) **N2NMN** uses hand-crafted and parameterized neural modules; (8) **LCGN** (Hu et al., 2019) uses a graph network where each node represents an object, and is described by a context-aware representation from related objects conditioned on the textual input.

To gain better insight into the relative contribution of the design choices we made, we perform experiment with the following ablated models: (9) **P-Ref+LSTM+Attn** uses attention instead of an external memory block for selecting the neighborhood words in the expression; (10) **P-Ref+Curriculum Learning**: We employ a curriculum training (Platanios et al. 2019) regime to train the P-Ref model in order to improve its performance without contextualization (See Appendix A.3).

5.3 Implementation Details

The memory matrix in our model discussed in section 4.2 consists of 128 rows and 80 columns. The controller is a single layer LSTM network. We use GloVe to obtain the word embedding (dimension = 300) of each word in the textual input. When training, we first train our program generator (PG) and use it as a fixed module for training the execution engine (EE). We use 18K ground-truth programs to train the program generator (PG). We train PG and EE using Adam (Kingma and Ba, 2015) with learning rates 0.0005 and 0.0001, respectively. Note that PG is trained for a maximum of 32,000 iterations, while EE is trained for a maximum of 450,000 iterations. We employ early stopping based on validation set accuracy. We do not find any significant improvements with the joint optimization of PG and EE. We train on one RTX 2080ti GPU with a batch size of 8.

5.4 Evaluation

Table 4 shows results in comparison with the baselines. We find that our contextual NMN model (P-Ref+LSTM+Mem) significantly outperforms all prior work by large margins. In addition to outperforming NMN baselines such as FiLM, N2NMN, IEP-Ref, we also outperform the non-NMN baselines such as LCGN demonstrating the effectiveness of the introduced memory module in captur-

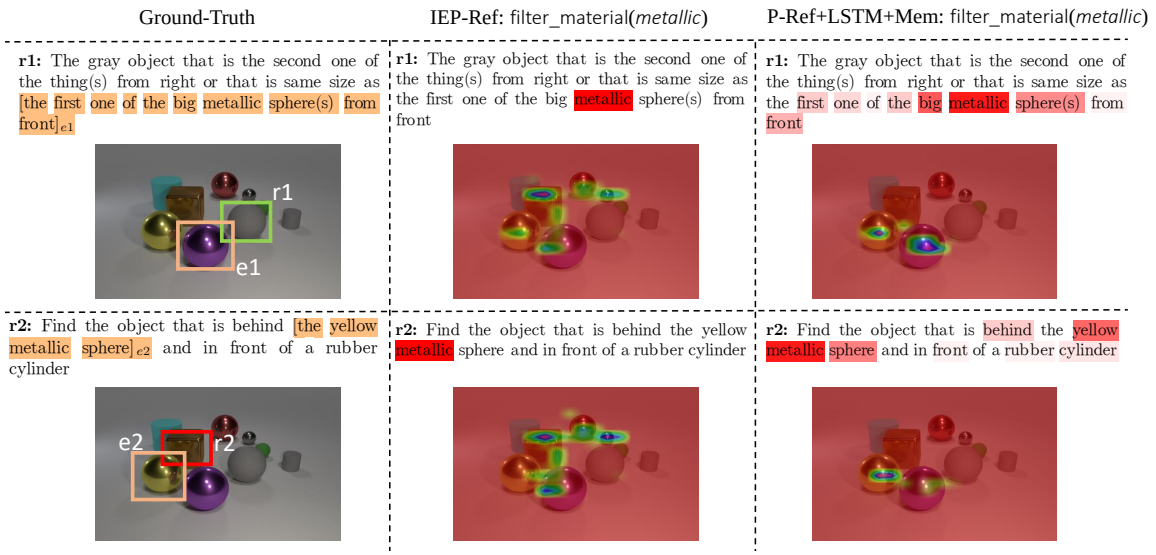


Figure 3: Qualitative examples showing the attention heatmaps of `filter_material(metallic)` module outputs trained using IEP-Ref and P-Ref+LSTM+Mem models. $e1$ and $e2$ highlight the metallic objects that are referred in the input expressions $r1$ and $r2$ respectively.

ing visiolinguistic relations and contextual dependencies from the longer CLEVR-like expressions. Specifically, we achieve +4.3% on F-test and +8.1% on S-Test, compared with the current state-of-the-art NMN model IEP-Ref. Most significant gains on S-Test also suggest the superior generalization skills of our model in learning from fewer training samples.

The ablation results are shown in Table 5. As we can see, all the ablative baselines underperform, confirming the importance of our proposed contextualization approach. Specifically the improvements obtained with module contextualization in both IEP-Ref and FiLM demonstrate that our approach can generalize across diverse NMN architectures.

Performance on CLOSURE and NLVR2 benchmarks is shown in Table 6. We achieve +11.2% in accuracy on CLOSURE test split compared to the best prior model Vector-NMN, indicating that our model generalizes well to unseen compositions. We also surpass all the existing NMN based models for NLVR2 dataset which has real images unlike synthetic images in CLEVR-Ref+ and CLOSURE.

Figure 3 illustrates the qualitative differences of `filter_material(metallic)` module trained using IEP-Ref and our P-Ref+LSTM+Mem model. With IEP-Ref, the model selects all metallic objects from the image, ignoring the context in the expression. On the other hand, our approach correctly locates objects based on their contextual relevance.

Model	F-Dev	F-Test	S-Dev	S-Test
IEP-Ref	80.54	78.20	49.89	51.50
FiLM	76.58	75.71	44.90	46.70
MAC	79.40	77.36	47.20	47.00
Vector-NMN	82.05	77.00	46.72	52.88
NS-VQA	80.08	79.01	48.07	51.66
N2NMN	76.00	75.11	43.62	46.70
LCGN	77.07	74.80	46.88	48.00
P-Ref+LSTM+Mem (**ours**)	**84.82**	**83.05**	**59.76**	**60.04**

Table 4: Performance of our memory based contextualized NMN model (P-Ref+LSTM+Mem) and baselines on CLEVR-Ref+.

5.4.1 Model Parameters

Our proposed model has 3 times fewer parameters than the baseline model IEF-Ref in total (see Table 7). More concretely, the baseline IEP-Ref model contains 60 modules and each module consists of 0.44M parameters. That is, total number of parameters in IEP-Ref are 60*0.44M = 26.4M. Similarly, the FiLM baseline, which also does contextualization of inputs, has 60*0.59M = 35.4M parameters. On the other hand, our proposed memory based contextualization of NMN model contains only a maximum of 15 modules and each module has 0.58M parameters. Therefore total number of parameters in our model are 15*0.58M = 8.7M. This is 3 times smaller than IEP-Ref and 4 times

Referring Expressions	500
Unique Images	492
Vocabulary	86
Avg. Length of Expr	20.4

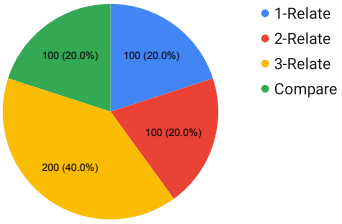


Figure 4: CC-Ref+ Statistics

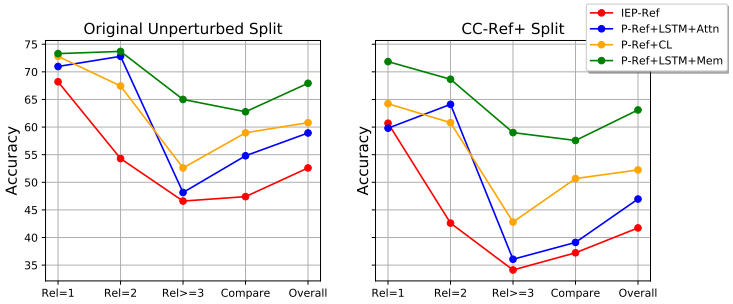


Figure 5: Performance of models on randomly drawn 500 original CLEVR-Ref+ test instances and their contrast sets.

Model	F-Dev	F-Test	S-Dev	S-Test
IEP-Ref	80.54	78.20	49.89	51.50
P-Ref+LSTM+Attn	79.26	78.99	52.68	52.96
P-Ref+LSTM+Mem (**ours**)	**84.82**	**83.05**	**59.76**	**60.04**
FiLM	76.58	75.71	44.90	46.70
FiLM+LSTM+Mem (**ours**)	79.05	80.86	51.10	53.06
P-Ref+CL	81.70	80.32	57.25	56.91
P-Ref+LSTM+Mem+CL	82.16	80.93	57.90	58.14

Table 5: **Ablations.** Performance of our model and its ablative baselines on CLEVR-Ref+.

Model	CLOSURE	NLVR2 (Test-P)
IEP-Ref	59.80	N/A
FiLM	58.72	51.10
N2NMN	62.07	52.10
MAC	65.19	51.40
Vector-NMN	64.14	N/A
P-Ref	59.68	N/A
P-Ref+LSTM+Attn	63.13	N/A
P-Ref+LSTM+Mem (**ours**)	**71.22**	N/A
FiLM+LSTM+Mem (**ours**)	69.78	**53.80**

Table 6: Performance of our model and NMN baselines on CLOSURE and NLVR2 datasets.

smaller than FiLM.

5.5 The CC-Ref+ Dataset

We further examine the robustness of the models by creating contrast sets (similar to Gardner et al. 2020) that help in exposing model brittleness by probing a model's decision boundary local to examples in the test set. Specifically, we follow a three stage approach to collect our contrast set:

Stage 1: First, we randomly sample 100 single-referent expressions from the test split containing only a single spatial relation (e.g. *The first one of the tiny rubber thing from **left***). We then sample another 100 expressions containing two spatial relations (e.g. *The first one of the thing from **left** that is **behind** the big yellow matte object*). Similarly we sample a third subset of 200 expressions containing 3 or more relations. Finally, we sample 100 expressions containing at least one compare relations (e.g. *Any other tiny object as the **same color** as the big yellow metallic cube*). This constitutes a total of 500 expressions.

Stage 2: We then manually perturb the semantics of various parts of these 500 referring expressions such that the ground-truth referent object changes. For example, we modify the expression *first one of the tiny rubber thing from left* to *first one of the tiny metallic thing from right*. We call this perturbed test split CC-Ref+. We show random selection of CC-Ref+ examples in Table 8.

Stage 3: Finally, we verify and validate the correctness of the new ground-truth annotations using two human annotators. The annotations that are not consistent among the two human annotators are removed and we re-iterate the above three steps until we collect a validated set of 500 contrast samples[3]. In Figure 4, we summarize the size and complexity of our CC-Ref+ split.

Model	#Parameters (per module)
IEP-Ref	442,752
Param. IEP-Ref (P-Ref)	574,336
FiLM	590,720
P-Ref+LSTM+Attn	574,464
P-Ref+CL	574,336
P-Ref+LSTM+Mem (**ours**)	589,597

Table 7: Count of parameters for each neural module in the baselines and our proposed NMN models.

[3](Gardner et al., 2020) shows that a few hundreds of contrast samples will be sufficient to draw substantiated conclusions about model behavior.

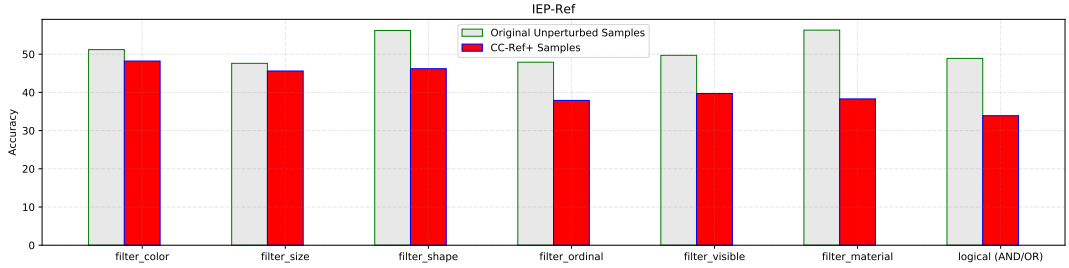


Figure 6: Performance of baseline IEP-Ref model on original CLEVR-Ref+ test split and CC-Ref+ samples

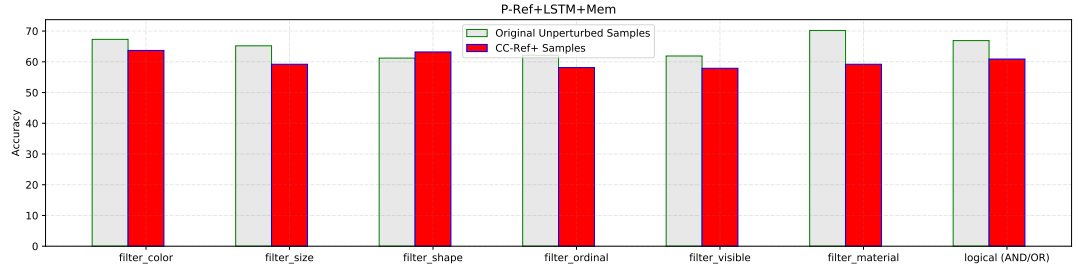
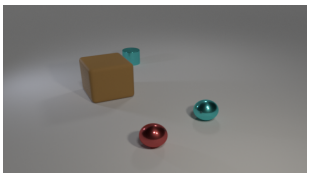
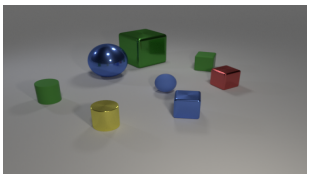


Figure 7: Performance of our contextual NMN model (P-Ref+LSTM+Mem) on original CLEVR-Ref+ test split and CC-Ref+ samples



Original: The brown things that are big object(s) or the second one of the small metal thing(s) from left
CC-Ref+: The cyan things that are big object(s) or the first one of the small metal thing(s) from left



Original: The matte things that are either the sixth one of the tiny thing(s) from right or the fifth one of the thing(s) from front
CC-Ref+: The matte things that are tiny thing(s) and the second one of the thing(s) from front

Table 8: Random examples from CC-Ref+ and their original annotations in CLEVR-Ref+

5.6 Evaluation on CC-Ref+ Dataset

As shown in Figure 5, performance of baseline models drop by >10% on CC-Ref+ and the models struggle to correctly ground the perturbed samples containing compare relations (e.g. *same_color*) or that containing more than 2 spatial relations (e.g. *front*, *left*) in the expression. Our method shows least drop (<5%) in performance indicating its superiority in grounding expressions with complex linguistic constructs (see Appendix B for more detailed analysis). In Figure 6 and Figure 7, we further analyze the model's performance when one of the object attributes namely, color, size, shape, material, ordinality, and visibility are perturbed in the contrast sets. We found that both IEP-Ref and our model are robust to perturbations in color indicating that this is a relatively easier concept to ground in the images. In contrast to the findings in (Liu et al., 2019), we see a significant drop by up to 15% in the performance of IEP-Ref on all the other attributes such as shape and visibility. Our proposed approach P-Ref+LSTM+Mem shows relatively low drop in the logical, material and ordinal perturbations, insignificant drops (< 3%) in color, visible perturbations and a slight improvement (+2%) in shape perturbations. This clearly suggests that our approach generalizes well and is robust to contrastive perturbations in the input. The performance gap of P-Ref+LSTM+Mem in logical, ordinal and material perturbations show that these

are relatively difficult concepts for the model to learn. We hope that CC-Ref+ dataset will foster more research in this area.

6 Conclusion

Neural module networks (NMNs) are widely used in language and vision tasks. We show that contextualizing these modules dramatically reduces the number of modules required and improve their grounding abilities, achieving a new state-of-the-art results on the CLEVR-Ref+ visual referring expressions task. Our analysis on CLEVR-Ref+, CLOSURE, NLVR2 and a new contrast set CC-Ref+ demonstrate that our proposed method enhances NMNs' ability to exploit visiolinguistic relationships.

Acknowledgements

We would like to thank Joyce Chai, Runtao Liu, Chenxi Liu and Yutong Bai for helpful discussions. We are grateful to the anonymous reviewers for their useful feedback.

References

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020a. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics.

Arjun R Akula. 2015. A novel approach towards building a generic, portable and contextual nlidb system. *International Institute of Information Technology Hyderabad*.

Arjun R. Akula, Shuai Wang, and Song-Chun Zhu. 2020b. Cocox: Generating conceptual and counterfactual explanations via fault-lines. 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 2594–2601. AAAI Press.

Arjun R Akula and Song-Chun Zhu. 2019. Visual discourse parsing. *ArXiv preprint*, abs/1903.02252.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. 2019. Closure: Assessing systematic generalization of clevr models. *ArXiv preprint*, abs/1912.05783.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6756–6764. AAAI Press.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets. *ArXiv preprint*, abs/2004.02709.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *ArXiv preprint*, abs/1410.5401.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017a. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 804–813. IEEE Computer Society.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10293–10302. IEEE.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017b. Modeling relationships in referential expressions with compositional modular networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition,*

CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 4418–4427. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2019. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5901–5914.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Changsong Liu, Shaohua Yang, Sari Saba-Sadiya, Nishant Shukla, Yunzhong He, Song-Chun Zhu, and Joyce Chai. 2016. Jointly learning grounded task structures from language instruction and visual demonstration. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1492, Austin, Texas. Association for Computational Linguistics.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Ashish Palakurthi, Ruthu S M, Arjun Akula, and Radhika Mamidi. 2015. Classification of attributes in a natural language query into different SQL clauses. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 497–506, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Hang Qi, Tianfu Wu, Mun-Wai Lee, and Song-Chun Zhu. 2015. A restricted visual turing test for deep scene and event understanding. *ArXiv preprint*, abs/1512.01715.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. REVERIE: remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 9979–9988. IEEE.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2014. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159, San Diego, California. Association for Computational Linguistics.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1039–1050.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, Sofia, Bulgaria. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

A Appendix

In this supplementary material, we begin by providing more details on CLEVR-Ref+ F-Ref / S-Ref splits and the neural modules in IEP-Ref to supplement Section 2 and Section 3 of the main paper, respectively. We then provide the details of our models (e.g., initialization & training, hyperparameters). Finally, we provide CC-Ref+ dataset annotation details, statistics, random examples, and more analysis to supplement Section 4 of the main paper.

A.1 F-Ref and S-Ref splits in CLEVR-Ref+

Visual referring expression recognition is the task of identifying the object in an image that is referred to by a natural language expression (Kazemzadeh et al., 2014; Mao et al., 2016). It is a fundamental language-to-vision matching problem and has several downstream applications such as question answering (Zhu et al., 2016). CLEVR-Ref+ (Liu et al., 2019) is a recently proposed dataset for visual referring expression recognition (RefExp) task, which consists of synthetic images and referring expressions. Specifically, it contains the ground-truth functional program representations that describe the intermediate visual reasoning as a chain of logical operations (i.e., neural modules) that need to be executed to find the target referent object (e.g., filter color, compare, filter size, and relate). There are nearly 0.8M referring expressions of which 32% of expressions refer to only a single object (*Single-referent*) and 68% refer to more than one object (*Multi-referent*). In this paper, we refer to the full dataset as F-Ref and the single-referent subset as S-Ref. Detailed statistics of the splits are presented in Table 9.

		F-Ref	S-Ref
Train	#Expr.	628915	200313 (32% of F-Ref)
	#Images	70000	62016
Dev	#Expr.	69879	22256
	#Images	6500	5200
Test	#Expr.	149741	47731
	#Images	15000	13534

Table 9: Statistics of F-Ref and S-Ref.

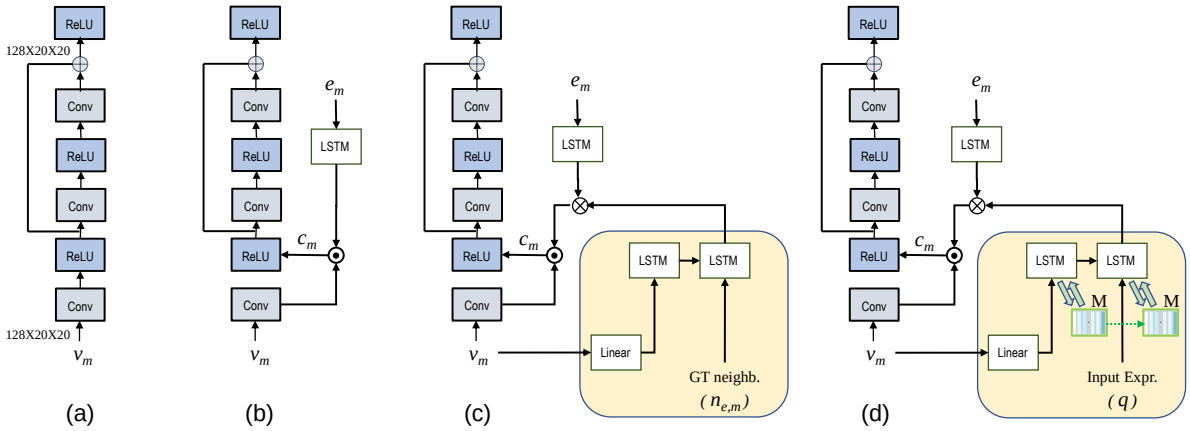


Figure 8: (a) Architecture of neural module (m) in IEP-Ref consuming a visual input v_m. $\oplus$ denotes summation. (b) Our proposed module design with parameterized textual input e_m. $\odot$ denotes element-wise multiplication. (c) Contextualized module design using ground-truth annotations for constructing neighborhood context ($n_{e,m}$) of e_m. $\otimes$ denotes concatenation. (d) Contextualized module design using our proposed memory (M) based architecture that learns to select the most relevant neighborhood context directly from the input expression q.

A.2 Neural Modules in Parameterized IEP-Ref

IEP-Ref (Liu et al., 2019), the current state-of-the-art neural module network (NMN) model for the CLEVR-Ref+ dataset, uses a generic design of neural module architecture adapted from IEP (Johnson et al., 2017b), which was designed for VQA task.[4] The modules take either two visual inputs (binary modules) or one visual input (unary modules). There are total 60 distinct modules in IEP-Ref. After parameterization (see Figure 8b), the distinct number of modules drop to 15 without any drop in the model performance (section 2 of main paper). That is, the number of a distinct set of modules (and the total number of parameters) used in the parameterized model reduces by 75%. Moreover, although the network parameters of each parameterized module slightly increase due to the additional LSTM unit, since each module in IEP-Ref can have multiple instantiations for the same textual input, we have fewer parameters than IEF-Ref in total. Table 11 presents the list of all the 15 modules in our parameterized NMN model. We compare the parameters per module of all baseline NMN models and our proposed models (section 3 of main paper) in Table 10.

Note that our proposed model has 3 times fewer parameters than the baseline model IEF-Ref in total. More concretely, the baseline IEP-Ref model contains 60 modules and each module consists of 0.44M parameters. That is, total number of parameters in IEP-Ref are 60*0.44M = 26.4M. Similarly, the FiLM baseline, which also does contextualization of inputs, has 60*0.59M = 35.4M parameters. On the other hand, our proposed memory based contextualization of NMN model contains only a maximum of 15 modules and each module has 0.58M parameters. Therefore total number of parameters in our model are 15*0.58M = 8.7M. This is 3 times smaller than IEP-Ref and 4 times smaller than FiLM.

[4]We used the IEP-Ref implementation provided at the link https://github.com/ruotianluo/iep-ref

Model	#Parameters (per module)
IEP-Ref	442,752
Param. IEP-Ref (P-Ref)	574,336
FiLM	590,720
P-Ref+LSTM+Attn	574,464
P-Ref+CL	574,336
P-Ref+LSTM+Mem	589,597

Table 10: Count of parameters for each neural module in the baselines and our proposed NMN models.

A.3 Model and other Experiment Details

Our proposed model (LSTM+Mem): The memory matrix consists of 128 rows and 80 columns. The controller is a single layer LSTM network. We use GloVe to obtain the word embedding (dimension = 300) of each word in the textual input. When training, we first train our program generator (PG) and use it as a fixed module for training the execution engine (EE).

	Modules
Unary	`Filter_Shape`, `Filter_Color`, `Filter_Material`, `Filter_Visible`, `Filter_Size`, `Filter_Ordinal`, `Unique`, `Relate`, `Same_Size`, `Same_Shape`, `Same_Color`, `Same_Material`, `Scene`
Binary	`Intersect`, `Union`

Table 11: Modules in Parameterized IEP-Ref

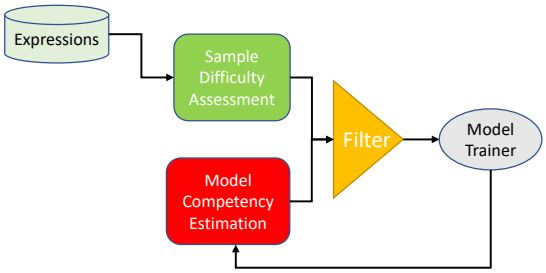


Figure 9: Overview of our curriculum learning baseline.

We use 18K ground-truth programs to train the program generator (PG). We train PG and EE using Adam (Kingma and Ba, 2015) with learning rates 0.0005 and 0.0001, respectively. Note that PG is trained for a maximum of 32,000 iterations, while EE is trained for a maximum of 450,000 iterations. We employ early stopping based on validation set accuracy. We do not find any significant improvements with the joint optimization of PG and EE. We train on one RTX 2080ti GPU with a batch size of 8.

Curriculum Learning Baseline: Prior literature shows that curriculum learning (CL) may greatly facilitate the learning of complex tasks for neural architectures (Platanios et al., 2019). Therefore, we employ a curriculum training (CL) regime as an additional baseline to train the P-Ref model in order to improve its performance without contextualization. An overview of the CL model is shown in Figure 9. To estimate the difficulty of the expressions, we define a scoring function inspired by what we, as humans, intuitively may consider difficult when grounding the expressions:

- Longer expressions are difficult to ground.
- Expressions with a large number of spatial relationships such as “left", “front", “right", “behind" are more likely to have difficult linguistic structures.
- Expressions requiring a large number of neural modules are difficult to ground.
- Expressions involving comparison modules are difficult to ground.

Using the above heuristics, we evaluate the difficulty of all expressions in the training set on a scale of 1 to 10. During the training, we initialize the model competency to 1. All the training expressions with difficulty level less than or equal to the current model competency are used for training the model. We use a validation set of expressions for each of these difficulty levels. As the model’s performance on the validation set starts to saturate, we increment the competency level of the model. We stop training immediately after the model’s competency reaches above 10. We use GloVe to obtain the word embedding (dimension = 300) of each word in the textual input. When training, we first train our program generator (PG) and use it as a fixed module for training the execution engine (EE). We use 18K ground-truth programs to train the program generator (PG). We train PG and EE using Adam (Kingma and Ba, 2015) with learning rates 0.0005 and 0.0001, respectively. PG is trained for a maximum of 32,000 iterations, and EE is trained for a maximum of 450,000 iterations. We employ early stopping based on validation set accuracy. We do not observe any significant improvements with the joint optimization of PG and EE. All of our CL experiments were conducted on one RTX 2080ti GPU with a batch size of 8.

B CC-Ref+ Annotation, Statistics, and Visualization

Following Gardner et al. 2020, we construct a contrast set for CLEVR-Ref+ dataset to identify systematic gaps (e.g., annotation artifacts) in the test split, and we call it CC-Ref+. Contrast sets help in exposing model brittleness by probing a model’s decision boundary local to examples in the test set.

Referring Expressions	500
Unique Images	492
Vocabulary	86
Expressions with #Relations $= 1$	100
Expressions with #Relations $= 2$	100
Expressions with #Relations ≥ 3	200
Expressions with *Compare* Relation	100
Avg. Length of Expression	20.2

Table 12: CC-Ref+ Statistics

We follow a three stage approach to collect our contrast set:

Stage 1: First, we randomly sample 100 single-referent expressions from the test split containing only a single spatial relation (e.g. *The first one of the tiny rubber thing from **left***). We then sample another 100 expressions containing two spatial relations (e.g. *The first one of the thing from **left** that is **behind** the big yellow matte object*). Similarly we sample a third subset of 200 expressions containing 3 or more relations. Finally, we sample 100 expressions containing at least one compare relations (e.g. *Any other tiny object as the **same color** as the big yellow metallic cube*). This constitutes a total of 500 expressions.

Stage 2: We then manually perturb the semantics of various parts of these 500 referring expressions such that the ground-truth referent object changes. For example, we modify the expression *first one of the tiny rubber thing from left* to *first one of the tiny metallic thing from right*. We show random selection of CC-Ref+ examples in Table 13.

Stage 3: Finally, we verify and validate the correctness of the new ground-truth annotations using two human annotators. The annotations that are not consistent among the two human annotators are removed and we re-iterate the above three steps until we collect a validated set of 500 contrast samples[5]. In Table 12, we summarize the size and complexity of our CC-Ref+ split.

B.1 Detailed Analysis of Models on CC-Ref+

In section 4.2 of main paper, we compared the performance of baseline models and our proposed method on CC-Ref+ in terms of number of relations (e.g. *in the front*, *to the left*, *of same shape as*) present in the expressions. In this section, we present more analysis in terms of object attributes. In CLEVR-Ref+, there are six types of object attributes namely, color, size, shape, material, ordinality, and visibility. We analyze the model's performance when one of these attributes are perturbed in the contrast sets. Additionally, we also compare the performance on contrast examples that involve logical AND/OR modifications. An example of contrast sample in CC-Ref+ involving logical AND/OR perturbation is as follows:

Original: The objects that are either the first one of the small metal object(s) from right *or* the first one of the metallic cube(s) from left.

CC-Ref+: The objects that are first one of the small rubber object(s) from right *and* the first one of the metallic object from front.

Figure 10 shows the performance of baseline IEP-Ref model on original test split and CC-Ref+ samples using the above attributes. Similarly, Figure 11, Figure 12, and Figure 13 shows the performance of models P-Ref+LSTM+Attn, P-Ref+CL, and P-Ref+LSTM+Mem respectively. We found that all the four models are robust to perturbations in color indicating that this is a relatively easier concept to ground in the images. In contrast to the findings in (Liu et al., 2019), we see a significant drop by up to 15% in the performance of baseline models on all the other attributes such as shape and visibility. P-Ref+CL also experience significant drops in accuracy on CC-Ref+. However it is found to be relatively more robust to the perturbations compared to the other baselines indicating that curriculum learning helps in adapting to contrast sets. Our proposed approach P-Ref+LSTM+Mem shows relatively low drop in the logical, material and ordinal perturbations, insignificant drops ($< 3\%$) in color, visible perturbations and a slight improvement (+2%) in shape perturbations. This clearly suggests that our approach generalizes well and is robust to perturbations in the input. The performance gap of P-Ref+LSTM+Mem in logical, ordinal and material perturbations show that these are relatively difficult concepts for the model to learn. We hope that CC-Ref+ dataset will foster more research in this area.

[5](Gardner et al., 2020) shows that a few hundreds of contrast samples will be sufficient to draw substantiated conclusions about model behavior.

References

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020a. Words aren't enough, their order matters: On the robustness of

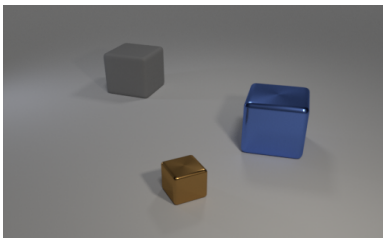
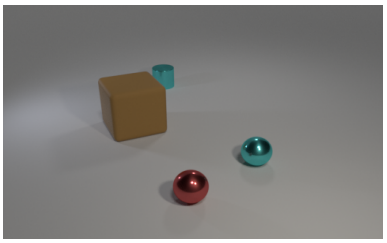
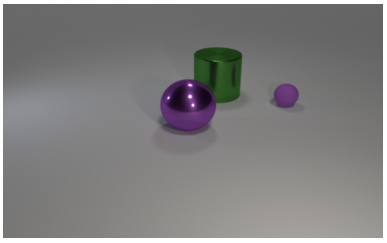
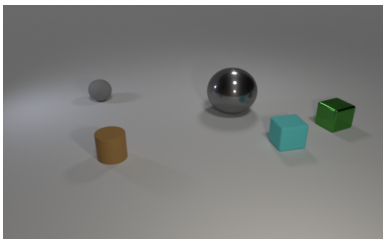
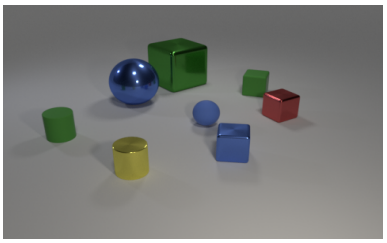
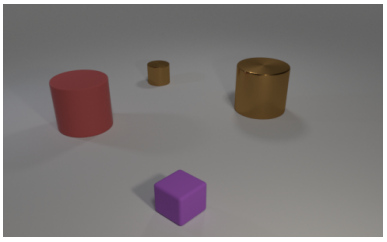
	**Original:** The big objects that are the first one of the block(s) from right or metallic object(s) **CC-Ref+:** The big objects that are the first one of the block(s) from left and rubber object(s)
	**Original:** The brown things that are big object(s) or the second one of the small metal thing(s) from left **CC-Ref+:** The cyan things that are big object(s) or the first one of the small metal thing(s) from left
	**Original:** The small objects that are the third one of the object(s) from left or purple shiny ball(s) **CC-Ref+:** The large objects that are the third one of the object(s) from left or purple shiny ball(s)
	**Original:** The tiny things that are the first one of the sphere(s) from left or the fourth one of the object(s) from front **CC-Ref+:** The tiny things that are the first one of the cylinder(s) from left or the fourth one of the object(s) from front
	**Original:** The matte things that are either the sixth one of the tiny thing(s) from right or the fifth one of the thing(s) from front **CC-Ref+:** The matte things that are tiny thing(s) and the second one of the thing(s) from front
	**Original:** The things that are either object(s) that are behind the tiny brown rubber thing(s) or the first one of the tiny brown thing(s) from left **CC-Ref+:** The things that are either object(s) that is in front of the tiny brown metallic thing(s) or the second one of the tiny brown thing(s) from left

Table 13: Random examples from CC-Ref+ and their original annotations in CLEVR-Ref+

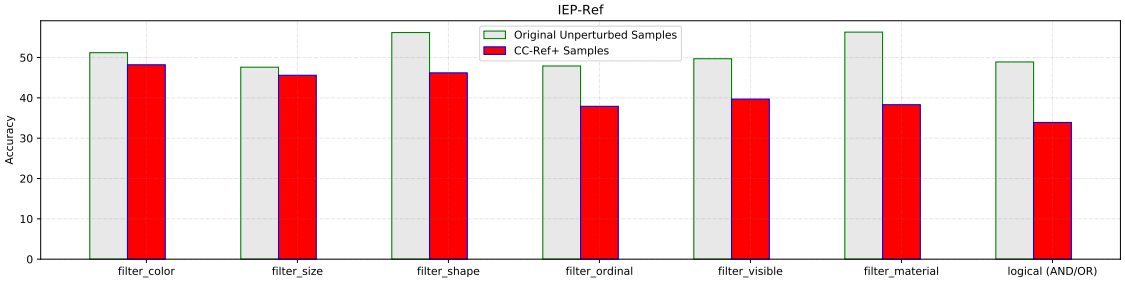


Figure 10: Performance of baseline IEP-Ref model on original test split and CC-Ref+ samples

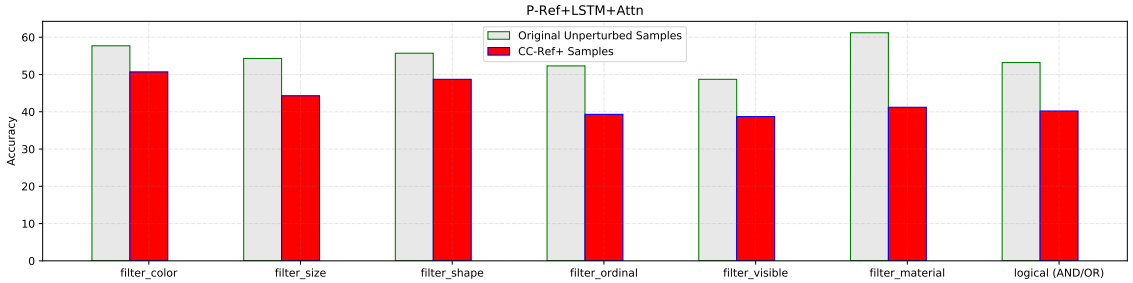


Figure 11: Performance of baseline P-Ref+LSTM+Attn model on original test split and CC-Ref+ samples

grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics.

Arjun R Akula. 2015. A novel approach towards building a generic, portable and contextual nlidb system. *International Institute of Information Technology Hyderabad*.

Arjun R. Akula, Shuai Wang, and Song-Chun Zhu. 2020b. Cocox: Generating conceptual and counterfactual explanations via fault-lines. 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 2594–2601. AAAI Press.

Arjun R Akula and Song-Chun Zhu. 2019. Visual discourse parsing. *ArXiv preprint*, abs/1903.02252.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, San Diego, California. Association for Computational Linguistics.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Dzmitry Bahdanau, Harm de Vries, Timothy J O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron Courville. 2019. Closure: Assessing systematic generalization of clevr models. *ArXiv preprint*, abs/1912.05783.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6756–6764. AAAI Press.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets. *ArXiv preprint*, abs/2004.02709.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *ArXiv preprint*, abs/1410.5401.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017a. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy,*

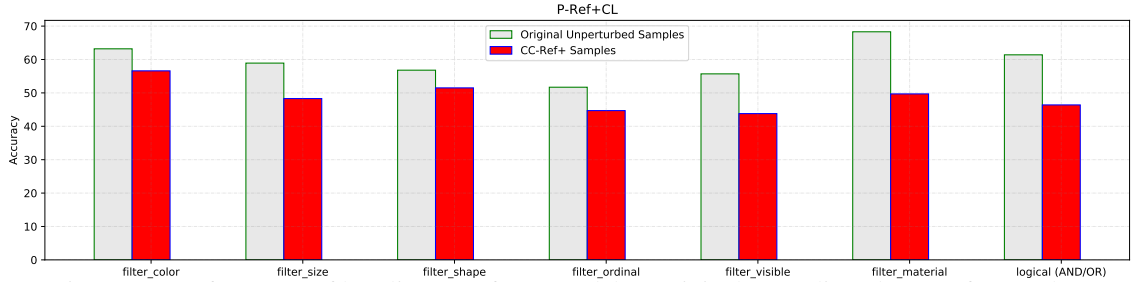


Figure 12: Performance of baseline P-Ref+CL model on original test split and CC-Ref+ samples.

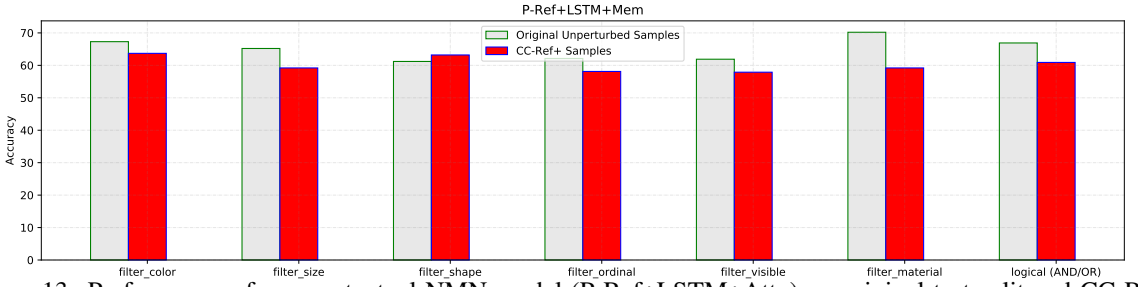


Figure 13: Performance of our contextual NMN model (P-Ref+LSTM+Attn) on original test split and CC-Ref+ samples

October 22-29, 2017, pages 804–813. IEEE Computer Society.

Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10293–10302. IEEE.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017b. Modeling relationships in referential expressions with compositional modular networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4418–4427. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2019. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5901–5914.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Changsong Liu, Shaohua Yang, Sari Saba-Sadiya, Nishant Shukla, Yunzhong He, Song-Chun Zhu, and Joyce Chai. 2016. Jointly learning grounded task structures from language instruction and visual

demonstration. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1492, Austin, Texas. Association for Computational Linguistics.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Ashish Palakurthi, Ruthu S M, Arjun Akula, and Radhika Mamidi. 2015. Classification of attributes in a natural language query into different SQL clauses. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 497–506, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Hang Qi, Tianfu Wu, Mun-Wai Lee, and Song-Chun Zhu. 2015. A restricted visual turing test for deep scene and event understanding. *ArXiv preprint*, abs/1512.01715.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. REVERIE: remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9979–9988. IEEE.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2014. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Y. Chai. 2016. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159, San Diego, California. Association for Computational Linguistics.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1039–1050.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, Sofia, Bulgaria. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2016, Las Vegas, NV, USA, June 27-30, 2016, pages
4995–5004. IEEE Computer Society.