CLEVR_HYP: A Challenge Dataset and Baselines for Visual Question Answering with Hypothetical Actions over Images

Shailaja Keyur Sampat[*], **Akshay Kumar, Yezhou Yang and Chitta Baral**
Arizona State Universiy, USA
{ssampa17,akuma216,yz.yang, chitta}@asu.edu

Abstract

Most existing research on visual question answering (VQA) is limited to information explicitly present in an image or a video. In this paper, we take visual understanding to a higher level where systems are challenged to answer questions that involve mentally simulating the hypothetical consequences of performing specific actions in a given scenario. Towards that end, we formulate a vision-language question answering task based on the CLEVR (Johnson et al., 2017a) dataset. Wethen modify the best existing VQA methods and propose baseline solvers for this task. Finally, we motivate the development of better vision-language models by providing insights about the capability of diverse architectures to perform joint reasoning over image-text modality[1].

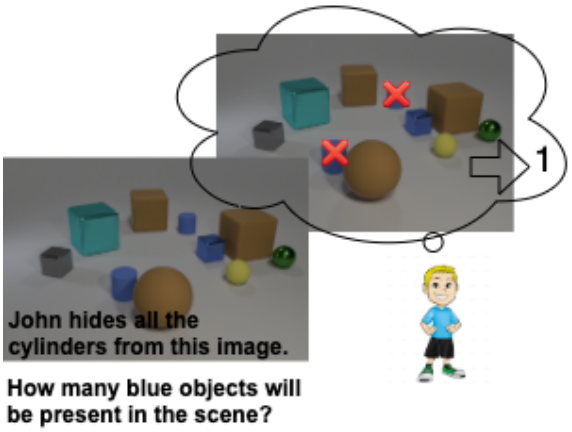


Figure 1: Motivation for the proposed CLEVR_HYP dataset: an example demonstrating how humans can do mental simulations and reason over resulting scenario.

1 Introduction

In 2014, Michael Jordan, in an interview (Gomes, 2014) said that "*Deep learning is good at certain problems like image classification and identifying objects in the scene, but it struggles to talk about how those objects relate to each other, or how a person/robot would interact with those objects. For example, humans can deal with inferences about the scene: what if I sit down on that?, what if I put something on top of something? etc. There exists a range of problems that are far beyond the capability of today's machines.*"

While this interview was six years ago, and since then there has been a lot of progress in deep learning and its applications to visual understanding. Additionally, a large body of visual question answering (VQA) datasets (Antol et al., 2015; Ren et al., 2015; Hudson and Manning, 2019) have been compiled and many models have been developed over them, but the above mentioned "inferences about the scene" issue stated by Jordan remains largely unaddressed.

In most existing VQA datasets, scene understanding is holistic and questions are centered around information explicitly present in the image (i.e. objects, attributes and actions). As a result, advanced object detection and scene graph techniques have been quite successful in achieving good performance over these datasets. However, provided an image, humans can speculate a wide range of implicit information. For example, the purpose of various objects in a scene, speculation about events that might have happened before, consider numerous imaginary situations and predicting possible future outcomes, intentions of a subject to perform particular actions, and many more.

Among the above, an ability to imagine taking specific actions and simulating probable results without actually acting or experiencing is an important aspect of human cognition (Figure 1 gives an example of this). Thus, we believe that having autonomous systems equipped with a similar capability will further advance AI research. This is particu-

[*] corresponding author

[1] Dataset setup scripts and code for baselines are made available at https://github.com/shailaja183/clevr_hyp. For additional details about the dataset creation process, refer supplementary material.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3692–3709
June 6–11, 2021. ©2021 Association for Computational Linguistics

larly useful for robots performing on-demand tasks in safety-critical situations or navigating through dynamic environments, where they imagine possible outcomes for various situations without executing instructions directly.

Motivated by the above, we propose a challenge that attempts to bridge the gap between state-of-the-art AI and human-level cognition. The main contributions of this paper[2] are as follows;

- We formalize a novel question answering task with respect to a hypothetical state of the world (in a visual form) when some action (described in a textual form) is performed.
- We create a large-scale dataset for this task, and refer it as `CLEVR_HYP` i.e. VQA with hypothetical actions performed over images in CLEVR (Johnson et al., 2017a) style.
- We first evaluate the direct extensions of top VQA and NLQA (Natural language QA) solvers on this dataset. Then, we propose new baselines to solve `CLEVR_HYP` and report their results.
- Through analysis and ablations, we provide insights about the capability of diverse architectures to perform joint reasoning over image-text modality.

2 Related Work

In this section we situate and compare our work with related areas such as implicit text generation/retrieval for a visual, visual question answering (VQA) over synthetic images, question answering (QA) involving hypothetical reasoning, and language-based manipulation in visual domains closest to `CLEVR_HYP`.

Implicit Text Generation for a Visual: VisualComet (Park et al., 2020) and Video2Commonsense (Fang et al., 2020) have made initial attempts to derive implicit information about images/videos contrary to traditional factual descriptions which leverage only visual attributes. VisualComet aims to generate commonsense inferences about events that could have happened before, events that can happen after and people's intents at present for each subject in a given image. They use a vision-language transformer that takes a sequence of inputs (image, event, place, inference) and train a model to predict inference in a language-model style. Video2Commonsense focuses on generating video descriptions that can incorporate commonsense facts related to intentions, effects, and implicit attributes about actions being performed by a subject. They extract top-ranked commonsense texts from the Atomic dataset and modify training objective to incorporate this information.

While both involve a visual-textual component and actions, their key focus is about generating plausible events and commonsense respectively. Whereas, our work is related to performing certain actions and reasoning about its effect on the overall visual scene.

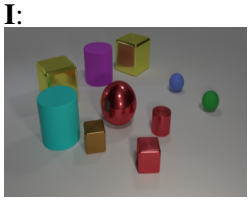
Language-based Manipulation in Visual Domain: Learning a mapping from natural language instructions to a sequences of actions to be performed in a visual environment is a common task in robotics (Kanu et al., 2020; Gaddy and Klein, 2019; Shridhar et al., 2020). Another relevant task is vision-and-language navigation (Anderson et al., 2018; Chen et al., 2019; Nguyen et al., 2019), where an agent navigates in a visual environment to find goal location by following natural language instructions. Both above works include visuals, natural language instructions and a set of actions that can be performed to achieve desired goals. In this way, it is similar to our `CLEVR_HYP`, but in our case, models require reasoning about the effect of actions performed rather than determining which action to perform. Also, we frame this in a QA style evaluation rather than producing instructions for low-level controls.

Manipulation of natural images with language is an emerging research direction in computer vision. (Teney et al., 2020) proposed a method for generating counterfactual of VQA samples using image in-painting and masking. Also, there are works (Dong et al., 2017; Nam et al., 2018; Reed et al., 2016) which use Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) for language conditioned image generation and manipulation. However, both the above tasks are more focused at object and attribute level manipulation rather than at action level.

VQA over Synthetic Images: While natural images-based VQA datasets reflect challenges one can encounter in real-life situations, the require-

[2]Our work focuses on the capability of neural models to reason about the effects of actions given a visual-linguistic context and not on models that deal with intuitive physics.

I:



1. $\mathbf{T}_A$: Paint the small green ball with cyan color.
   $\mathbf{Q}_H$: Are there equal yellow cubes on left of purple object and cyan spheres? (A: yes)
2. $\mathbf{T}_A$: Add a brown rubber cube behind the blue sphere that inherits its size from the green object.
   $\mathbf{Q}_H$: How many things are either brown or small? (A: 6)
3. $\mathbf{T}_A$: John moves the small red cylinder on the large cube that is to the right of purple cylinder.
   $\mathbf{Q}_H$: What color is the object that is at the bottom of the small red cylinder? (A: yellow)

Figure 2: Three examples from CLEVR_HYP dataset: given image (I), action text (T_A), question about hypothetical scenario (Q_H) and corresponding answer (A). The task is to understand possible perturbations in I with respect to various action(s) performed as described in T_A. Questions test various reasoning capabilities of a model with respect to the results of those action(s).

ment of costlier human annotations and vulnerability to biases are two major drawbacks. Contrary to them, synthetic datasets allow controlled data generation at scale while being flexible to test specific reasoning skills.

For the above reasons, following benchmark VQA datasets have incorporated synthetic images; COG (Yang et al., 2018) and Shapes (Andreas et al., 2016) contain images with rendered 2D shapes; SHRDLU (Winograd, 1971), CLEVR (Johnson et al., 2017a), and CLEVR-dialog (Kottur et al., 2019) have rendered scenes with 3D objects; DVQA (Kafle et al., 2018) and FigureQA (Kahou et al., 2017) have synthetically generated charts (bar chart, pie chart, dot-line etc.); VQA-abstract (Antol et al., 2015) and IQA (Gordon et al., 2018) involves question-answering over synthetically rendered clipart-style scenes and interactive environments respectively. Our proposed dataset CLEVR_HYP uses CLEVR (Johnson et al., 2017a) style rendered scenes with 3D objects as a visual component. It is distinct from all other synthetic VQA datasets for two key reasons; first, integration of action domain in synthetic VQA and second, the requirement of mental simulation in order to answer the question.

QA involving Hypothetical Reasoning: In the language domain, WIQA (Tandon et al., 2019) dataset tests the model's ability to do what-if reasoning over procedural text as a 3-way classification (the influence between pair of events as positive, negative or no-effect). In vision-language domains, a portion of TQA (Kembhavi et al., 2017) and VCR (Zellers et al., 2019) are relevant. Questions in TQA and VCR involve hypothetical scenarios about multi-modal science contexts and movie scenes respectively. However, none of the above two datasets' key focus is on the model's capability to imagine changes performed over the image.

As shown in Figure 3, the setting of TIWIQ (a benchmark dataset for "physical intelligence") (Wagner et al., 2018) has some similarity with ours. It has synthetically rendered table-top scenes, four types of actions (push, rotate, remove and drop) being performed on an object and what-if questions.

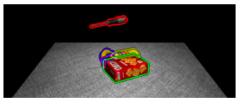
Input Scene And Action	Output Description
 1) "the robot drops the screw driver on the foam" — what happens to the foam?	**Ground Truth** "the foam is pushed a little by the screw driver" **Human Baseline** "the foam is pushed because the screw driver drops on it" **Hybrid Model Prediction** "the foam is pushed by the screw driver" **Data-Driven Model Prediction** "the foam is pushed by the mustard container"

Figure 3: Example from TIWIQ (Wagner et al., 2018).

To our best knowledge, TIWIQ dataset is not publicly available. Based on our understanding from their manuscript, we observe following important distinction with this work. Our questions focus on the impact of actions on the whole image, while in TIWIQ questions are about impact of actions on a specific object in the image. Moreover, we frame CLEVR_HYP as a classification task, contrary to TIWIQ which is a generative task. Our CLEVR_HYP dataset has 175k automatically generated image-action text-question samples which is much larger compared to TIWIQ which has only 1020 samples and manually crafted ground-truths.

3 CLEVR_HYP Task and Dataset

Figure 2 gives a glimpse of CLEVR_HYP task. We opt for synthetic dataset creation as it allows automated and controlled data generation at scale with minimal biases. More details are described below.

3 Inputs: Image(I), Action Text (T_A) and Hypothetical Question (Q_H)

1. Image(I): It is a given visual for our task. Each image in the dataset contains 4-10 randomly selected 3D objects rendered using Blender (Blender Online Community, 2019) in CLEVR

(Johnson et al., 2017a) style. Objects have 4 attributes listed in the Table 1. Additionally, these objects can be referred using 5 relative spatial relations (left, right, in front, behind and on). We provide scene graphs[3] containing all ground-truth information about a scene, that can be considered as a visual oracle for a given image.

Attr.	Possible values in CLEVR_HYP
Color	gray, blue, brown, yellow, red, green, purple, cyan
Shape	cylinder, sphere or cube
Size	small or big
Material	metal (shining) or rubber (matte)

Table 1: Object attributes in CLEVR_HYP scenes.

2. Action Text (T_A): It is a natural language text describing various actions performed over the current scene. The action can be one of four:

(i) **Add** new object(s) to the scene
(ii) **Remove** object(s) from the scene
(iii) **Change** attributes of the object(s)
(iv) **Move** object(s) within scene (might be in plane i.e. left/right/front/back or out of plane i.e. move one object on top of another object[4])

To generate action text, we start with manually written templates involving the aforementioned actions. For example, action involving change in the attribute of object(s) to a given value, we have a template of the following kind; 'Change the <A> of <Z><C><M><S> to <V>'. Where <A>, <Z>, <C>,<M>,<S>,<V> are placeholders for the attribute, size, color, material, shape and a value of attribute respectively. Each action text in the CLEVR_HYP is associated with a functional program which if executed on an image's scene graph, yields the new scene graph that simulates the effects of actions.

Functional programs for action texts[3] are built from the basic functions that correspond to elementary action operations (right part of Figure 4a). For the above mentioned 'change' attribute action template, the equivalent functional program can be written as; 'change_attr(<A>,filter_size(<Z>,filter_color(<C>, filter_material(<M>filter_shape(<S>, scene()))))),<V>)'. It essentially means, first filter out the objects with desired attributes and then update the value of their current attribute A to value V.

3. Question about Hypothetical Situation (Q_H): It is a natural language query that tests various visual reasoning abilities after simulating the effects of actions described in T_A. There are 5 possible reasoning types similar to CLEVR;

(i) **Counting** objects fulfilling the condition
(ii) **Verify existence** of certain objects
(iii) **Query attribute** of a particular object
(iv) **Compare attributes** of two objects
(v) **Integer comparison** of two object sets (same, larger or smaller)

Similar to action texts, we have templates and corresponding programs for questions. Functional programs for questions[3] are executed on the image's updated scene graph (after incorporating effects of the action text) and yields the ground-truth answer to the question. Functional programs for questions are made of primitive functions shown in left part of the Figure 4a).

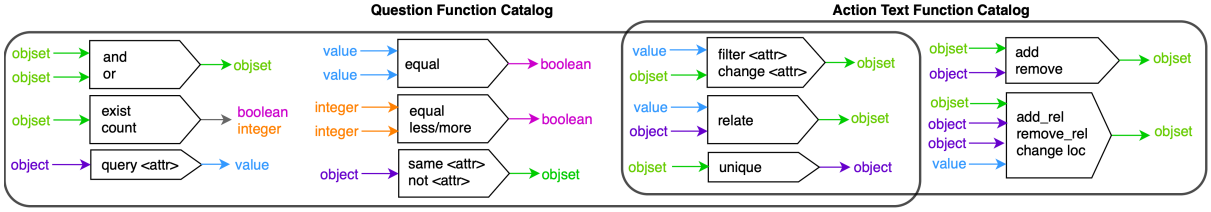
Paraphrasing: In order to create a challenging dataset from linguistic point of view and to prevent models from overfitting on templated representations, we leverage noun synonyms, object name paraphrasing and sentence-level paraphrasing. For noun synonyms, we use a pre-defined dictionary (such as cubeb̃lock, sphereb̃all and so on). We programmatically generate all possibilities to refer to an object in the image (i.e. object name paraphrasing) and randomly sample one among them. For sentence level paraphrasing, we use Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020) fine-tuned over positive samples from Quora Question Pairs (QQP) dataset (Iyer et al., 2017) for question paraphrasing. We use Fairseq (Ott et al., 2019) for action text paraphrasing which uses round-trip translation and mixture of experts (Shen et al., 2019).

Note that we keep the action text and question as separate inputs for the purpose of simplicity and keeping our focus on building solvers that can do mental simulation. One can create a simple template like "<Q_H> if <proper-noun/pronoun> <T_A>?" or "If <proper-noun/pronoun> <T_A>,

[3]Scene graphs and Functional Programs (for action text and question) are not provided at the test-time.

[4]For simplicity, we assume that any object can be put on another object regardless of its size, material or shape.

(a) Function Catalog for CLEVR_HYP, extended from CLEVR (Johnson et al., 2017a)



(b) Dataset creation pipeline

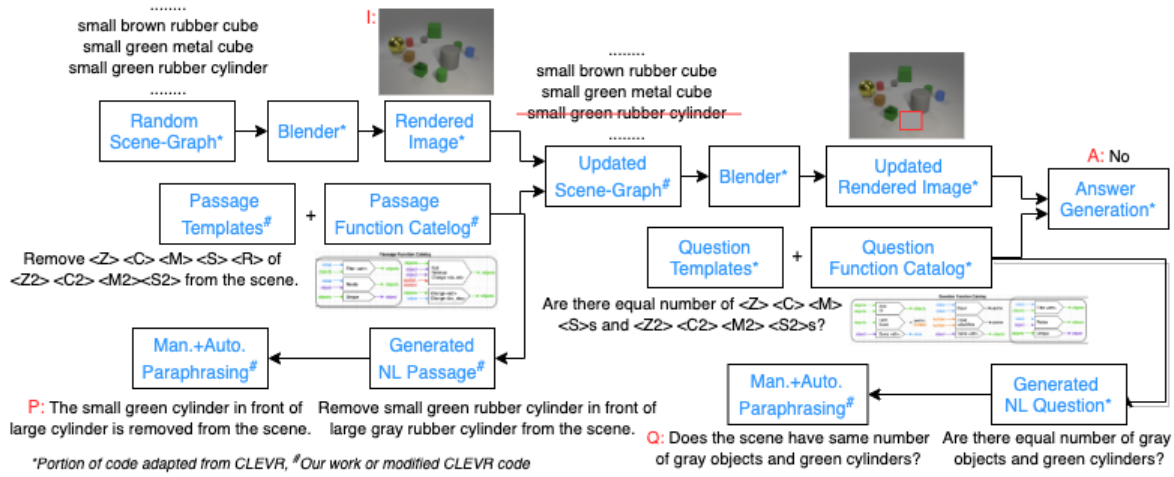


Figure 4: CLEVR_HYP dataset creation process with example and function catalog used for ground-truth answer generation. (for more details, see Appendix A.4)

<Q_H>?" if they wish to process action and question as a single text input. For example, "How many things are the same size as the cyan cylinder if I add a large brown rubber cube behind the blue object." or "If I add a large brown rubber cube behind the blue object, how many things are the same size as the cyan cylinder?". However, having them together adds further complexity on the solver side as it first has to figure out what actions are performed and what is the question.

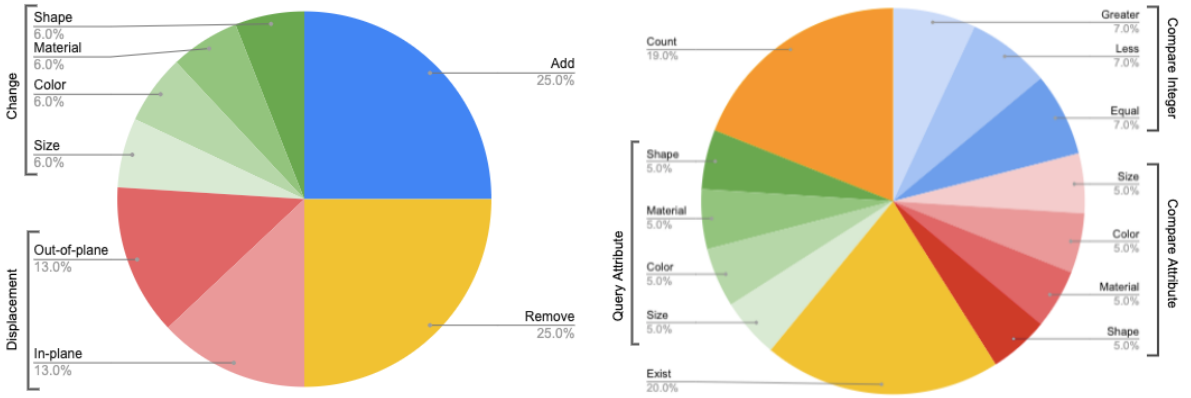
By providing ground-truth object information (as a visual oracle) and machine-readable form of questions & action texts (oracle for linguistic components). This information can be used to develop models which can process semi-structured representations of image/text or for the explainability purposes (to precisely know which component of the model is failing).

Output: **Answer (A)** to the Question (Q_H), which can be considered as a 27-way classification over attributes (8 colors + 3 shapes + 2 sizes + 2 material), numeric (0-9) and boolean (yes/no).

Dataset Partitions and Statistics: We create CLEVR_HYP dataset containing 175k image-action text-question samples using the process mentioned in Figure 4b. For each image, we generate 5 kinds of action texts (one for each add, remove, move in-plane and move out-of-plane and change attribute). For each action text type, we generate 5 questions (one for each count, exist, compare integer, query attribute and compare attribute). Hence, we get 5*5 unique action text-question pairs for each image, covering all actions and reasoning types in a balanced manner as shown in Figure 5a (referred as Original partition). However, it leads to a skewed distribution of answers as observed from 5b. Therefore, we curate a version of the dataset (referred as Balanced partition) consisting of 67.5k samples where all answer choices are equally-likely as well.

Additionally, we create two small challenge test sets (1500 image-action text-question samples each)- 2HopActionText (2HopT$_A$) and 2HopQuestion (2HopQ$_H$) to test generalization capability of the trained models. In 2HopT$_A$, we create action text which requires model to understand two different actions being taken on the scene. For example, 'Add a small blue metal cylinder to the right of large yellow cube and remove the large cylinder from the scene.' and 'Move the purple object on top of small red cube then change its color to cyan.'. In

(a) Distribution based on Action Text types and Question types



(b) Distribution of Answer types

Figure 5: Visualization of distributions for actions, questions and answers in Original_Train partition of CLEVR_HYP.

2HopQ$_H$, we create questions which require model to understand logical combinations of questions using 'and', 'or' and 'not'. For example, 'How many objects are either red or cylinder?' and 'Are there any rubber cubes that are not green?'.

In Table 2, we provide size of the various partitions and measure the diversity of the dataset in various aspects. For images, we calculate average number of objects present in the scene from the length of scene graph. For balanced partition, the number of images are much less compared to original, but more average number of objects per image. This is most likely due to the need to accommodate integers 4-9 more frequently as ground-truth answers. For textual components, we show average lengths (number of tokens separated by whitespaces) and count unique utterances as a measure of diversity. The original partition of the resulting dataset has 80% and 83% unique action text and questions respectively. For balanced partition, length and unique utterances for action text are nearly same as the original partition but for questions, it decreases. Questions in the original partition have been observed to enforce more strict and specific object references (such as small red metal cubes) compared to balanced partition (small cubes, red metal objects etc.), reducing the average length and uniqueness. It is intuitive for 2Hop partitions to have higher average length and uniqueness for T_A and Q_H respectively. This shows that despite having created this dataset from templates and rendered images with a limited set of attributes, it is still fairly challenging.

4 Models that we experiment with

Models trying to tackle CLEVR_HYP dataset have to address four key challenges;

(i) understand hypothetical actions and questions in complex natural language,
(ii) correctly disambiguate the objects of interest and obtain the structured representation (i.e. scene graphs or functional programs) of various modalities if required by the solver,
(iii) understand the dynamics of the world based on the various actions performed over it,
(iv) perform various kind of reasoning to answer the question.

4.1 Random

The QA task in CLEVR_HYP dataset can be considered as a 27-class classification problem. Each answer choice is likely to be picked with a probability of 1/27. Therefore, the performance of the random baseline is 3.7%.

Split	#I	Avg. #Obj	#T$_A$	Unique #T$_A$	Avg. T_A Len.	#Q$_H$	Unique #Q$_H$	Avg. Q$_H$ Len.
Original_Train	5k	6.4	25k	20.7k	12.8	125k	103.7k	22.6
Original_Val	1k	6.7	5k	3.8k	12.8	25k	20.9k	23.1
Original_Test	1k	6.4	5k	3.6k	12.6	25k	20.7k	22.8
Balanced_Train	5k	7.6	25k	21.1k	12.8	67.5k	58.2k	20.3
Balanced_Val	1k	7.6	5k	3.9k	12.7	13.5k	11.5k	20.7
Balanced_Test	1k	7.5	5k	3.7k	12.6	13.5k	11.4k	20.4
2HopT_A_Test	1k	6.4	3k	2.6k	18.6	15k	12.5k	22.8
2HopQ_H_Test	1k	6.4	3k	2.2k	12.6	15k	13.7k	29.3

Table 2: CLEVR_HYP dataset splits and statistics (# represents number of, k represents thousand).

4.2 Human Performance

We performed human evaluation with respect to 500 samples from the CLEVR_HYP dataset. Accuracy of human evaluations on original test, 2HopA_T and 2HopQ_H are 98.4%, 96.2% and 96.6% respectively.

4.3 Transformer Architectures

Pre-trained transformer-based architectures have been observed (Li et al., 2020) to capture a rich hierarchy of language-structures (text-only models) and effectively map entities/words with corresponding image regions (vision-language models). We experiment with various transformer-based models to understand their capability to understand the effects of actions on a visual domain.

Baseline 1- Machine Comprehension using RoBERTa: To evaluate the hypothetical VQA task through the text-only model, we convert images into the templated text using scene graphs. The templated text contains two kind of sentences; one describing properties of the objects i.e. "There is a <Z> <C> <M> <S>", the other one describing the relative spatial location i.e. "The <Z> <C> <M> <S> is <R> the <Z1> <C1> <M1> <S1>". For example, "There is a small green metal cube." and "The large yellow rubber sphere is to the left of the small green metal cube". Then we concatenate templated text with the action text to create a reading comprehension passage. We use state-of-the-art machine comprehension baseline RoBERTa (Liu et al., 2019) finetuned on the RACE dataset (Lai et al., 2017)[5]. Finally, we predict an answer to the question using this reading comprehension passage.

Baseline 2- Visual Question Answering using LXMERT Proposed by (Tan and Bansal, 2019), LXMERT is one of the best transformer based pre-trainable visual-linguistic representations which supports VQA as a downstream task. Typical VQA systems take an image and a language input. Therefore, to evaluate CLEVR_HYP in VQA style, we concatenate action text and question to form a single text input. Since LXMERT is pre-trained on the natural images, we finetune it over CLEVR_HYP dataset[6] and then use it to predict answer.

4.4 Systematically incorporating effects of actions into neural models

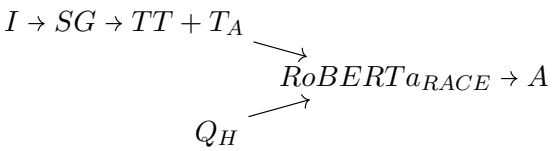
Baseline 3- Text-editing Image Baseline: In this method, we break-down the QA task with mental simulation in two parts; first, learn to generate an updated image (such that it has incorporated the effects of actions) and then perform visual question answering with respect to the updated image. We use the idea from Text Image Residual Gating proposed in (Vo et al., 2019) to implement the first part. However there are two important distinctions; Their focus is on the retrieval from the given database. We modify their objective and develop text-adaptive encoder-decoder with residual connections to generate new image. Also, editing instructions in their CSS dataset (Vo et al., 2019) were quite simple. For example, 'add red cube' and 'remove yellow sphere'. In this case, one can add the red cube anywhere in the scene. We modify their architecture to precisely place objects to their

[5]architecture=roberta large, epochs=5, learning rate=1e−05, batch size=2, update frequency=2, dropout=0.1, optimizer=adam with eps=1e−06.

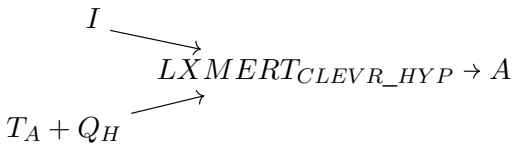
[6]epochs=4, learning rate=5e−05, batch size=8

Nomenclature I: Image, SG: Scene Graph, TT: Templated Text, T_A: Action Text, Q_H: Hypothetical Question, A: Answer, FP: Functional Program, ': Updated Modality

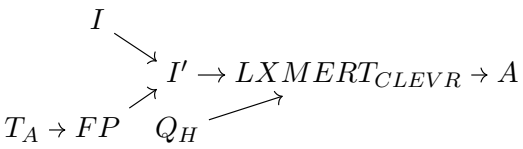
Baseline 1:



Baseline 2:



Baseline 3:



Baseline 4:

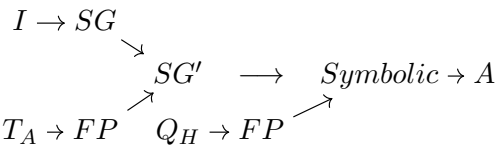


Figure 6: Graphical visualization of baseline models over CLEVR_HYP described above.

relative spatial references (on left/right/front/ behind). Once we get the updated image, we feed it to the LXMERT (Tan and Bansal, 2019) finetuned over the CLEVR (Johnson et al., 2017a) dataset along with the question and predict the answer.

Baseline 4- Scene Graph Update Model: Instead of directly manipulating images, in this method, we leverage image scene graphs to convert image-editing problem into graph-editing problem, conditioned on the action text. This is an emerging research direction to deal with changes in the visual modality over time or with new sources of information, as observed from recent parallel works (chang Chen et al., 2020; He et al., 2020).

We first use Mask R-CNN (He et al., 2017) to get the segmentation mask of the objects and predict attributes (color, material, size, and shape) with an acceptance threshold of 0.9. Segmentation mask of each object along with original image is then passed through ResNet-34 (He et al., 2016) to extract precise 3D coordinates of the object. We get the structured scene graph for the image. Then we use seq2seq with attention model originally proposed in (Johnson et al., 2017b) to generate functional programs (FP) for action text and question. The execution engine executes programs on scene graph, implemented as a neural module network (Andreas et al., 2016) to update the scene representation and answer questions.

We learn to update scene graphs according to functional program for the action text using reinforcement learning[7]. The reward function is associated with our ground-truth program executor and generates reward if prediction exactly matches with ground-truth execution. Once we get the updated scene representation, we use neural-symbolic model[8] proposed by (Yi et al., 2018) to obtain the final answer. It is notable that (Yi et al., 2018) achieved near-perfect performance on the CLEVR QA task in addition to being fully explainable.

5 Baseline Results

In this section, we benchmark models described above on the CLEVR_HYP. The dataset is formulated as a classification task with exactly one correct answer, so we use standard accuracy as evaluation metric. We then analyze their performance according to question and action types.

Quantitative results from above experiments can be visualized in top part of the Table 3. Among the methods described above, the scene graph update model has the best overall performance 70.5% on original test data. Text-editing model is best over balanced set, but observed to have the poor generalization capability when two actions or reasoning capabilities have to be performed. CLEVR_HYP requires models to reason about effect of hypothetical actions taken over images. LXMERT is not directly trained for this objective therefore, it struggles to do well on this task. The reason behind the poor performance of text-only baseline is due to its limitation to incorporate detailed spatial locations

[7] finetuning learning rate=1e−05, 1M iterations with early stopping, batch size=32

[8] supervised pretraining learning rate=7e−04, num iterations=20k, batch size=32 and then finetuning 1e−05, at most 2M iterations with early stopping, batch size=32

Overall Baseline Performance for Various Test Sets of CLEVR_HYP

Original Test				*Balanced Test*				*2HopTA Test*				*2HopQH Test*			
BL1	BL2	BL3	BL4	BL1	BL2	BL3	BL4	BL1	BL2	BL3	BL4	BL1	BL2	BL3	BL4
57.2	63.9	64.7	**70.5**	55.3	65.2	**69.5**	68.6	53.3	49.2	55.6	**64.4**	55.2	52.9	58.7	**66.5**

Performance break-down by Action Types and Reasoning Types for Baseline 3 and 4

	Original Test			*2HopA_T Test*			*Original Test*			*2HopQ_H Test*	
	BL3	BL4		BL3	BL4		BL3	BL4		BL3	BL4
Add	58.2	65.9	Add+Remove	53.6	63.2	Count	60.2	74.3	And	59.2	67.1
Remove	89.4	88.6	Add+Change	55.4	64.7	Exist	69.6	72.6	Or	58.8	67.4
Change	88.7	91.2	Add+Move	49.7	57.5	CompInt	56.7	67.3	Not	58.1	65.0
Move(in-plane)	61.5	69.4	Remove+Change	82.1	85.5	CompAttr	68.7	70.5			
Move(on)	53.3	66.1	Remove+Move	52.6	66.4	QueryAttr	65.4	68.1			
			Change+Move	53.8	63.3						

Table 3: Baseline performance over CLEVR_HYP (BLx represents one of the four Baselines described above).

into the templates that we use to convert image into a machine comprehension passage.

Two of our models (scene graph update and text-editing image) are transparent to visualize intermediate changes in the scene after performing actions. We analyse their ability to understand actions and make appropriate changes as shown in below part of Table 3. For the scene graph method, we compare the ground-truth functional program with the generated program and measure their exact-match accuracy. For the text-editing image method, we generate scene graphs for both images (original image and image after text-editing) and compare them. For attributes, we do exact-match, whereas for location information we consider matching only on the basis of relative spatial location.

Both scene graph and text-editing models do quite well on 'remove' and 'change' actions whereas struggle when new objects are added or existing objects are moved around. The observation is consistent when multiple actions are combined. Therefore, actions remove+change can be performed with maximum accuracy whereas other combinations of actions accomplish relatively lower performance. It leads to the conclusion that understanding the effect of different actions are of varied complexity. Most models demonstrate better performance over counting, existence and attribute query type of questions than comparison questions. The scene graph update and text-editing methods show a performance drop of 6.1% and 9.1% respectively when multiple actions are performed on the scene. However, there is less of a performance gap for models on 2HopQ_H compared to the test set, suggesting that models are able to better generalize with respect to multiple reasoning skills than complex actions.

6 Conclusion

We introduce CLEVR_HYP, a dataset to evaluate the ability of VQA systems after hypothetical actions are performed over the given image. We create this dataset by extending the data generation framework of CLEVR (Johnson et al., 2017a) that uses synthetically rendered images and templates for reasoning questions. Our dataset is challenging because rather than asking to reason about objects already present in the image, it asks about what would happen in an alternative world where changes have occurred. We provide ground-truth representations for images, hypothetical actions and questions to facilitate the development of models that systematically learn to reason about underlying process. We create several baseline models to benchmark CLEVR_HYP and report their results. Our analysis shows that the models are able to perform reasonably well (70.5%) on the limited number of actions and reasoning types, but struggle with complex scenarios. While neural models have achieved almost perfect performance on CLEVR and considering human performance as upperbound (98%), there is a lot of room for improvement on CLEVR_HYP. Our future work would include relaxing constraints by allowing a larger variety of actions, attributes and reasoning types. By extending this approach further for natural images, we aim to contribute in the development of better vision+language models.

Acknowledgements

We are thankful to the anonymous reviewers for the constructive feedback. This work is partially supported by the grants NSF 1816039, DARPA W911NF2020006 and ONR N00014-20-1-2332.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3674–3683. IEEE Computer Society.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. 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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Blender Online Community. 2019. *Blender - a 3D modelling and rendering package*. Blender Foundation.

Li chang Chen, Guosheng Lin, S. Wang, and Qingyao Wu. 2020. Graph edit distance reward: Learning to edit scene graph. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5707–5715. IEEE Computer Society.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860. Association for Computational Linguistics.

David Gaddy and Dan Klein. 2019. Pre-learning environment representations for data-efficient neural instruction following. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1956. Association for Computational Linguistics.

Lee Gomes. 2014. Machine-learning maestro michael jordan on the delusions of big data and other huge engineering efforts.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 4(5):6.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: visual question answering in interactive environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4089–4098. IEEE Computer Society.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Xuanli He, Quan Hung Tran, Gholamreza Haffari, Walter Chang, Zhe Lin, Trung Bui, Franck Dernoncourt, and Nhan Dam. 2020. Scene graph modification based on natural language commands. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 972–990. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *data. quora. com.*

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.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: understanding data visualizations via question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5648–5656. IEEE Computer Society.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

John Kanu, Eadom Dessalene, Xiaomin Lin, Cornelia Fermuller, and Yiannis Aloimonos. 2020. Following instructions by imagining and reaching visual goals. *arXiv preprint arXiv:2001.09373*.

Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5376–5384. IEEE Computer Society.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. 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 582–595. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: Manipulating images with natural language. 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 42–51.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12527–12537. Computer Vision Foundation / IEEE.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53. Association for Computational Linguistics.

J. Park, Chandra Bhagavatula, R. Mottaghi, A. Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1060–1069. JMLR.org.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. IEEE.

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. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 6076–6085. Association for Computational Linguistics.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Answering visual what-if questions: From actions to predicted scene descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 580–599.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6439–6448. Computer Vision Foundation / IEEE.

Misha Wagner, Hector Basevi, Rakshith Shetty, Wenbin Li, Mateusz Malinowski, Mario Fritz, and Ales Leonardis. 2018. Answering visual what-if questions: From actions to predicted scene descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, Massachusetts inst of tech cambridge project mac.

Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. 2018. A dataset and architecture for visual reasoning with a working memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 714–731.

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.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

A Appendix

A.1 Relation of CLEVR_HYP dataset with real-world situations

Teaching methodologies leverage our ability to mentally simulate scenarios along with the metaphors to aid understanding about new concepts. In other words, to explain unfamiliar concepts, we often reference familiar concepts and provide additional clues to establish mapping between them. This way, a person can create a mental simulation about unfamiliar concept and aid basic understanding about it.

For example, we want to explain a person how a 'zebra' looks like, who has previously seen a 'horse', we can do so using example in Figure 7a. This naturally follows for more complex concepts. Let say, one wants to describe the structure of an atom to someone, he might use the analogy of a planetary system, where the components (planets $\sim$ electrons) circulate around a central entity (sun $\sim$ nucleus). One more such example is provided in Figure 7b.

(a) learning the concept 'zebra' from the 'horse'

(b) learning about 'animal cell' by comparison with 'plant cell'

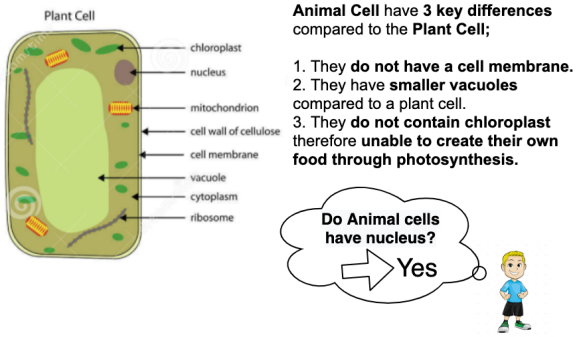


Figure 7: Extension of CLEVR_HYP for more complex real-world scenarios.

For humans, learning new concepts and performing mental simulations is omnipresent in day-to-day life. Therefore, CLEVR_HYP dataset is very much grounded in the real world. Models developed on this dataset can serve a broad range of applications, particularly the ones where possible outcomes have to be predicted without actually executing the actions. For example, robots performing on-demand tasks in safety-critical situations or self-driving vehicles. In addition, these models can be an important component for other vision and language tasks such as automatic expansion of existing knowledge bases, zero shot learning and spatio-temporal visual reasoning.

A.2 Rejecting Bad Samples in CLEVR_HYP

Automated methods of question generation sometimes create invalid items, classified as 'ill-posed' or 'degenerate' by CLEVR (Johnson et al., 2017a) dataset generation framework. They consider question "What color is the cube to the right of the sphere?" as ill-posed if there were many cubes right of the sphere, or degenerate if there is only one cube in the scene and reference to the sphere becomes unnecessary. In addition to this, we take one more step of quality control in order to prevent ordinary VQA models from succeeding over CLEVR_HYP without proper reasoning.

In CLEVR_HYP, one has to perform actions described in T over image I and then answer question Q with respect to the updated scenario. Therefore, to prevent ad-hoc models from exploiting biases in CLEVR_HYP, we pose the requirement that a question must have different ground-truth answers for CLEVR_HYP and image-only model. One such example is shown in Figure 8. For image (I), Q1 leads to different answers for CLEVR and CLEVR_HYP, making sure that one needs to correctly incorporate the effect of T. Q2 is invalid for a given image-action text pair in the CLEVR_HYP as one can answer it correctly without understanding T.

A.3 More Examples from CLEVR_HYP

Beyond Figure 10, all rest of the pages show more examples from our CLEVR_HYP dataset. Each dataset item has 4 main components- image(I), action text (T_A), question about the hypothetical states (Q_H) and answer (A). We classify samples based on what actions are taken over the image and the kind of reasoning is required to answer questions.

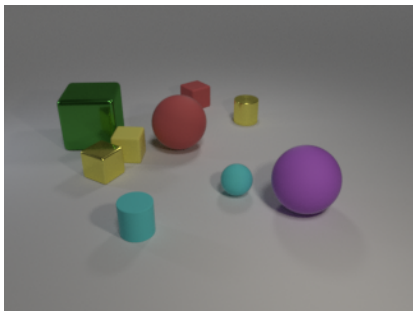
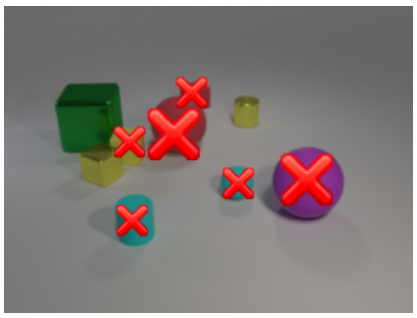
I: 

Image-only model:
Q1: Is there any large sphere? **A**: Yes
Q2: Is there any large cube? **A**: Yes

`CLEVR_HYP`:
T: Remove all matte objects from the scene.

I': 

Q1: Is there any large sphere? **A**: No ✓
Q2: Is there any large cube? **A**: Yes ✗

Figure 8: Validity of questions in `CLEVR_HYP`

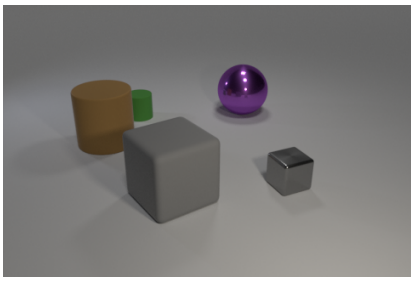
A.4 Function Catalog

As described in Section 3 and shown in Figure 4, each action text and question is associated with a functional program. We provide more details about these basic functions in Table 4 that was used to generate ground-truth answers for our dataset. Each function has input and output arguments, which are limited to following data types:

- object: a single object in the scene
- objset: a set of zero or more objects in scene
- integer: an integer in [0,10]
- boolean: 'yes' or 'no'
- values: possible attribute values mentioned in Table 1

A.5 Paraphrasing

In order to create a challenging dataset from the linguistic point of view and to prevent models from overfitting on templated representations, we leverage word synonyms and paraphrasing methods. This section provides more details about paraphrasing methods used in our dataset.



small gray metal cube: [small gray object, small metal object, small cube, small gray cube, small gray metal object, gray metal cube, small gray metal cube]

large brown rubber cylinder: [brown object, large brown object, large cylinder, brown rubber object, brown cylinder, large brown rubber object, large brown cylinder, brown rubber cylinder, large brown rubber cylinder]

Figure 9: Object paraphrases for 2 objects in the scene

Object Name Paraphrasing There can be many ways an object can be referred in the scene. For example, 'large purple metal sphere' in image below can also be referred to as 'sphere' as there is no other sphere present in the image. In order to make templates more challenging, we use these alternative expressions to refer objects in the action text or question. We wrote a python script that takes scene graph of the image and generates all possible names one can uniquely refer for each object in the scene. When paraphrasing is performed, one of the generated names is randomly chosen and replaced. Figure 9 demonstrates list of all possible name variants for two objects in the given image.

Synonyms for Paraphrasing We use word synonyms file provided with CLEVR dataset generation code.

Sentence/Question Level Paraphrasing For action text paraphrasing, we use Fairseq (Ott et al., 2019) based paraphrasing tool which uses round-trip translation and mixture of experts (Shen et al., 2019). Specifically, we use pre-trained round-trip models (En-Fr and Fr-En) and choose top-5 paraphrases manually for each template. For question paraphrasing, the quality of round-trip translation and mixture of experts was not satisfactory. Therefore, we use Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020) fine-tuned over positive samples from Quora Question Pairs (QQP) dataset (Iyer et al., 2017) and choose top-5 per template.

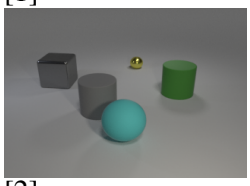
A.6 Computational Resources

All of our experiments are performed over Tesla V100-PCIE-16GB GPU.

Function	Input Type → Output Type	Return Value
scene	ϕ → objset	Set of all objects in the scene
unique	objset → object	Object if objset is singleton; else raise exception (to verify whether the input is unique or not)
relate	object × relation → objset	Objects satisfying given spatial relation for input object
count	objset → integer	Size of the input set
exist	objset → boolean	'Yes' if the input set is non-empty and 'No' otherwise
filter_size	objset × size → objset	Subset of input objects that match the given size
filter_color	objset × color → objset	Subset of input objects that match the given color
filter_material	objset × material → objset	Subset of input objects that match the given material
filter_shape	objset × shape → objset	Subset of input objects that match the given shape
query_size	object → size	Size of the input object
query_color	object → color	Color of the input object
query_material	object → material	Material of the input object
query_shape	object → shape	Shape of the input object
same_size	object → objset	Set of objects that have same size as input (excluded)
same_color	object → objset	Set of objects that have same color as input (excluded)
same_material	object → objset	Set of objects that have same material as input(excluded)
same_shape	object → objset	Set of objects that have same shape as input (excluded)
equal_size	size × size → boolean	'Yes' if inputs are equal, 'No' otherwise
equal_color	color × color → boolean	'Yes' if inputs are equal, 'No' otherwise
equal_material	material × material → boolean	'Yes' if inputs are equal, 'No' otherwise
equal_shape	shape × shape → boolean	'Yes' if inputs are equal, 'No' otherwise
equal_integer	integer × integer → boolean	'Yes' if two integer inputs are equal, 'No' otherwise
less_than	integer × integer → boolean	'Yes' if first integer is smaller than second, else 'No'
greater_than	integer × integer → boolean	'Yes' if first integer is larger than second, else 'No'
and	objset × objset → objset	Intersection of the two input sets
or	objset × objset → objset	Union of the two input sets.
not_size	object → objset	Subset of input objects that do not match given size
not_color	object → objset	Subset of input objects that do not match given color
not_material	object → objset	Subset of input objects that do not match given material
not_shape	object → objset	Subset of input objects that do not match given shape
add	objset × object → objset	Input set with input object added to it
remove	objset × object → objset	Input set with input object removed from it
add_rel	objset × object x object x relation → objset	Input set with new object (first input) added at the given spatial location relative to second input object
remove_rel	objset × object x object x relation → objset	Input set with object (first input) removed from the given spatial location relative to second input object
change_loc	objset × object x object x relation → objset	Input set with object (first input) location changed to a given spatial location relative to second input object
change_size	objset × size → objset	Input set with size updated to the given value
change_color	objset × color → objset	Input set with color updated to the given value
change_material	objset × material → objset	Input set with material updated to the given value
change_shape	objset × shape → objset	Input set with shape updated to the given value

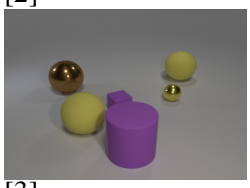
Table 4: (upper) Original function catalog for CLEVR proposed in (Johnson et al., 2017a), which we reuse in our data creation process (lower) New functions added to the function catalog for the CLEVR_HYP dataset.

[1]



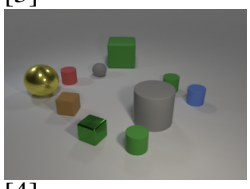
T$_A$: A small red sphere is added to the right of the green object.
Q$_H$: There is a gray cylinder; how many spheres are to the right of it?
A: 2
Classification: Add action, Counting question
Split: val

[2]



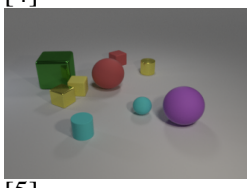
T$_A$: All the purple objects become metallic.
Q$_H$: What number of shiny things are to the left of the small yellow sphere?
A: 3
Classification: Change action, Counting question
Split: val

[3]



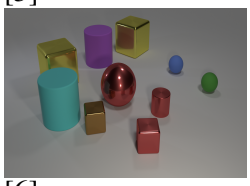
T$_A$: John puts a large red metal cube behind the blue rubber cylinder.
Q$_H$: There is a small green cylinder that is in front of the gray thing; are there any large red things behind it?
A: Yes
Classification: Add action, Existence question
Split: val

[4]



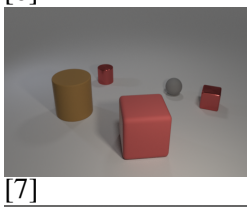
T$_A$: Remove all matte objects from the scene.
Q$_H$: Is there any large sphere?
A: No
Classification: Remove action, Existence question
Split: val

[5]



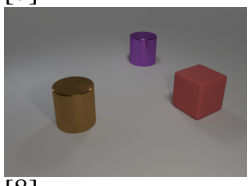
T$_A$: The large cylinder behind the red shiny sphere is moved in front of the green sphere.
Q$_H$: Is there a purple object that is to the right of the big yellow cube that is behind the cyan rubber sphere?
A: No
Classification: Move (in-plane) action, Existence question
Split: val

[6]



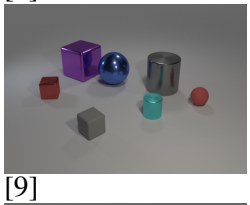
T$_A$: A small green metal sphere is added behind the small red cube.
Q$_H$: What color is the large cylinder that is to the right of the green object?
A: Brown
Classification: Add action, Query Attribute question
Split: val

[7]



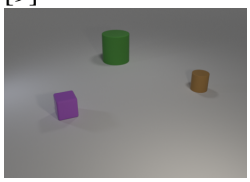
T$_A$: The purple cylinder behind the cube disappers from the scene.
Q$_H$: What material is the object on the left of brown metal cylinder?
A: Rubber
Classification: Remove action, Query Attribute question
Split: val

[8]



T$_A$: There is a sphere that is to the left of the gray cylinder; it shrinks in size.
Q$_H$: What size is the blue object?
A: Small
Classification: Change action, Query Attribute question
Split: val

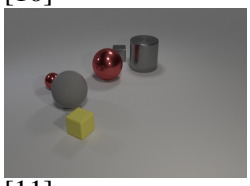
[9]



T$_A$: The brown thing is moved in front of the pink rubber cube.
Q$_H$: What shape is the object that is in front of the pink rubber cube?
A: Cylinder
Classification: Move (in-plane) action, Query Attribute question
Split: val

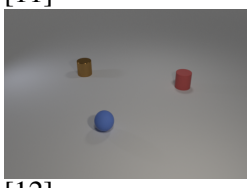
Figure 10: More examples from the CLEVR_HYP dataset

[10]



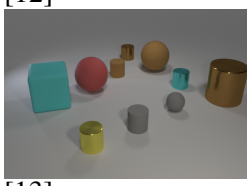
T$_A$: The small red sphere is moved onto the small cube that is in front of the gray sphere.
Q$_H$: What material is the object that is below the small metal sphere?
A: Rubber
Classification: Move (out-of-plane) action, Query Attribute question
Split: val

[11]



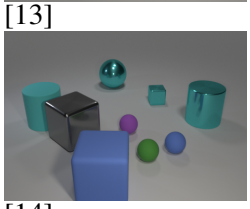
T$_A$: A small yellow metal object is placed to the right of red cylinder; it inherits its shape from the blue object.
Q$_H$: Are there any other things that have the same shape as the blue matte object?
A: Yes
Classification: Add action, Compare Attribute question
Split: val

[12]



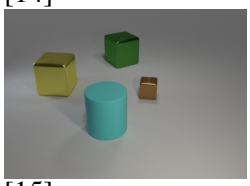
T$_A$: Hide all the cylinders from the scene.
Q$_H$: Are there any other things that have the same size as the gray sphere?
A: No
Classification: Remove action, Compare Attribute question
Split: val

[13]



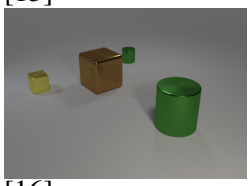
T$_A$: The small block is displaced and put on the left of the blue cube.
Q$_H$: Is there anything else on the right of the cyan sphere that has the same color as the large metal cylinder?
A: No
Classification: Move (in-plane) action, Compare Attribute question
Split: val

[14]



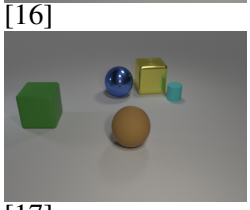
T$_A$: Jill places the small cube on the large cube that is to the left of cyan cylinder.
Q$_H$: There is an object below the brown cube; does it have the same shape as the green object?
A: Yes
Classification: Move (out-of-plane) action, Compare Attribute question
Split: val

[15]



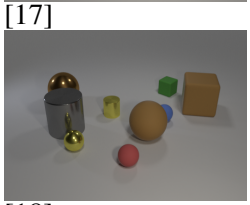
T$_A$: A small brown cube is added to the scene which is made of same material as the golden block.
Q$_H$: Are there an equal number of green objects and brown cubes?
A: Yes
Classification: Add action, Compare Integer question
Split: val

[16]



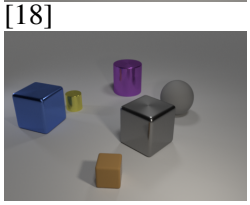
T$_A$: The tiny cylinder is withdrawn from the scene.
Q$_H$: Is the number of rubber objects greater than the number of shiny objects?
A: No
Classification: Remove action, Compare Integer question
Split: val

[17]



T$_A$: All small metal spheres are transformed into cylinders.
Q$_H$: Are there fewer brown objects that are to the right of the red sphere than the cylinders?
A: Yes
Classification: Change action, Compare Integer question
Split: val

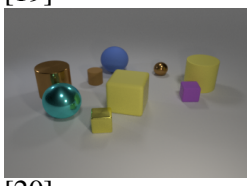
[18]



T$_A$: The sphere is placed in front of the large blue cube that is to the left of the yellow shiny object.
Q$_H$: Are there an equal number of gray things to the right of the brown rubber cube and cylinders?
A: No
Classification: Move (in-plane) action, Compare Integer question
Split: val

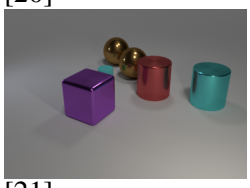
Figure 11: More examples from the CLEVR_HYP dataset

[19]



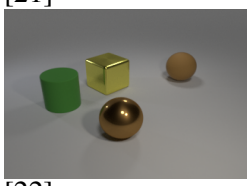
T$_A$: John hides the big object to the right of the brown sphere.
Q$_H$: How many yellow or cyan objects are there?
A: 3
Classification: Remove action, Counting question with 'Or'
Split: 2HopQ$_H$ test

[20]



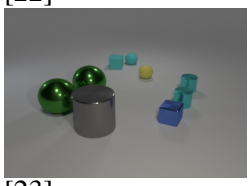
T$_A$: All brown things become matte.
Q$_H$: How many any other things are there which are made of the same material as the small cyan object?
A: 2
Classification: Change action, Counting + Compare Attribute question
Split: 2HopQ$_H$ test

[21]



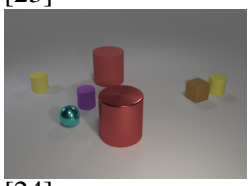
T$_A$: Make all the brown objects shiny.
Q$_H$: Are there any non metal things to the right of the shiny sphere?
A: No
Classification: Change action, Existence question with negation
Split: 2HopQ$_H$ test

[22]



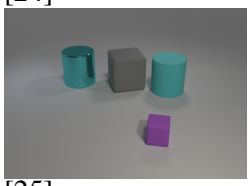
T$_A$: The gray object is moved to the right of the yellow thing.
Q$_H$: There is a cyan block; what number of big objects are there to the left of it that has the same material as the blue cube?
A: 2
Classification: Move (in-plane) action, Counting + Compare Attribute question
Split: 2HopQ$_H$ test

[23]



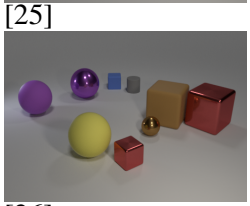
T$_A$: Remove all the yellow cylinders; Then clone the brown object and put it to the left side of the cyan ball.
Q$_H$: How many objects are made of the rubber?
A: 4
Classification: Add+Remove actions, Count question
Split: 2HopT$_A$ test

[24]



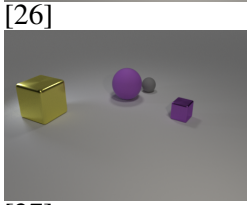
T$_A$: Enlarge the purple object; Then add a large red matte sphere to the right of the large purple cube.
Q$_H$: Is there any small object in the scene?
A: No
Classification: Add+Change actions, Existence question
Split: 2HopT$_A$ test

[25]



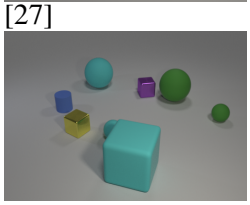
T$_A$: Add a small brown rubber sphere to the left of yellow matte object; Then swap its position with the purple shiny sphere.
Q$_H$: There is a ball that is to the left of the blue cube; what is its color?
A: Brown
Classification: Add+Move actions, Query Attribute question
Split: 2HopT$_A$ test

[26]



T$_A$: Sam takes the purple block out of the scene; Then he paints the yellow object by green color.
Q$_H$: Is there anything else that has the same material as the small gray sphere?
A: No
Classification: Remove+Change actions, Compare Attribute question
Split: 2HopT$_A$ test

[27]



T$_A$: Remove the cyan balls from the scene and move the large cyan cube on top of the yellow object.
Q$_H$: Are there greater number of spheres to the right of the yellow object than cubes?
A: No
Classification: Remove+Move actions, Compare Integer question
Split: 2HopT$_A$ test

Figure 12: More examples from the CLEVR_HYP dataset