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'

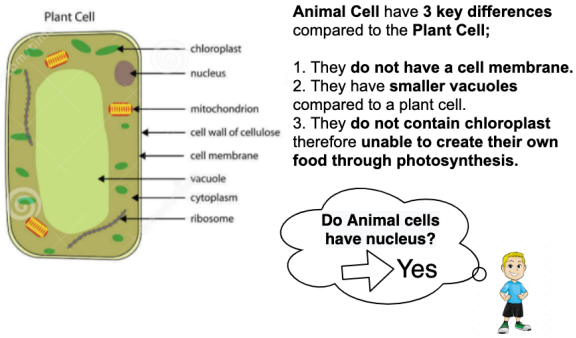


Figure 7: Extension of CLEVR_HYP for more complex real-world scenarios.

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.

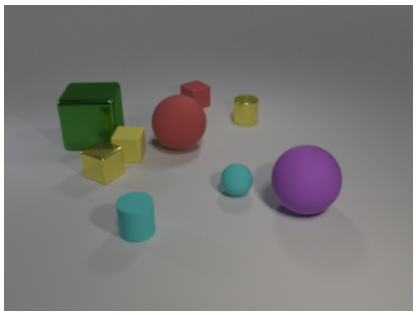
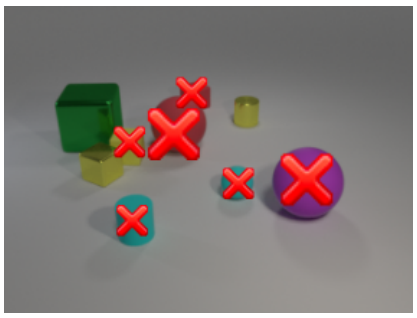
I: 

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`

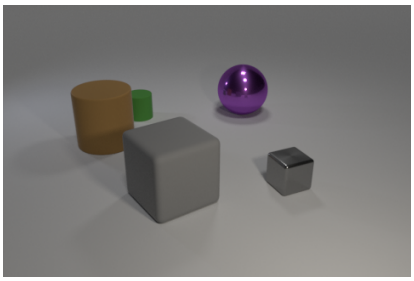
A.4 Function Catalog

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.

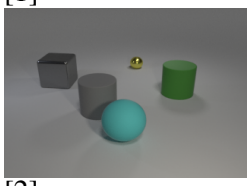
A.6 Computational Resources

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

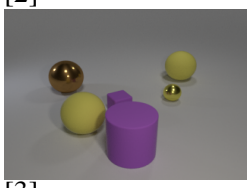
Table 4: (upper) Original function catalog for CLEVR proposed in (Johnson et al., 2017a), which we reuse in our data creation process (lower) New functions added to the function catalog for the CLEVR_HYP dataset.

[1]



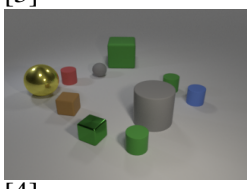
T$_A$: A small red sphere is added to the right of the green object.
Q$_H$: There is a gray cylinder; how many spheres are to the right of it?
A: 2
Classification: Add action, Counting question
Split: val

[2]



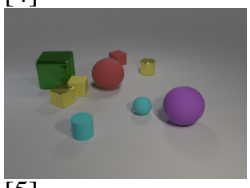
T$_A$: All the purple objects become metallic.
Q$_H$: What number of shiny things are to the left of the small yellow sphere?
A: 3
Classification: Change action, Counting question
Split: val

[3]



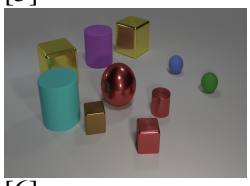
T$_A$: John puts a large red metal cube behind the blue rubber cylinder.
Q$_H$: There is a small green cylinder that is in front of the gray thing; are there any large red things behind it?
A: Yes
Classification: Add action, Existence question
Split: val

[4]



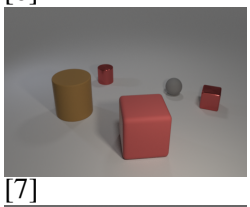
T$_A$: Remove all matte objects from the scene.
Q$_H$: Is there any large sphere?
A: No
Classification: Remove action, Existence question
Split: val

[5]



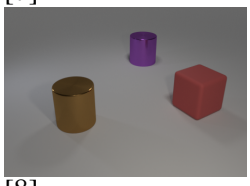
T$_A$: The large cylinder behind the red shiny sphere is moved in front of the green sphere.
Q$_H$: Is there a purple object that is to the right of the big yellow cube that is behind the cyan rubber sphere?
A: No
Classification: Move (in-plane) action, Existence question
Split: val

[6]



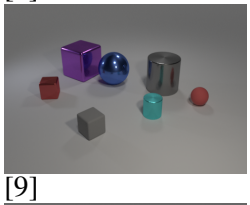
T$_A$: A small green metal sphere is added behind the small red cube.
Q$_H$: What color is the large cylinder that is to the right of the green object?
A: Brown
Classification: Add action, Query Attribute question
Split: val

[7]



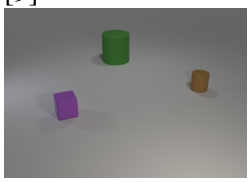
T$_A$: The purple cylinder behind the cube disappers from the scene.
Q$_H$: What material is the object on the left of brown metal cylinder?
A: Rubber
Classification: Remove action, Query Attribute question
Split: val

[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

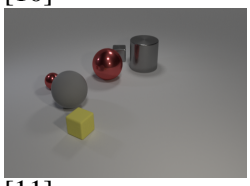
[9]



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

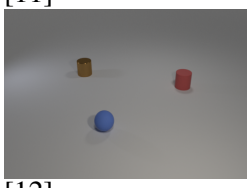
Figure 10: More examples from the CLEVR_HYP dataset

[10]



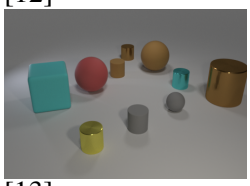
T$_A$: The small red sphere is moved onto the small cube that is in front of the gray sphere.
Q$_H$: What material is the object that is below the small metal sphere?
A: Rubber
Classification: Move (out-of-plane) action, Query Attribute question
Split: val

[11]



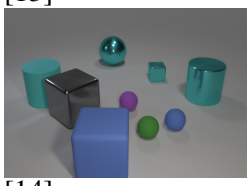
T$_A$: A small yellow metal object is placed to the right of red cylinder; it inherits its shape from the blue object.
Q$_H$: Are there any other things that have the same shape as the blue matte object?
A: Yes
Classification: Add action, Compare Attribute question
Split: val

[12]



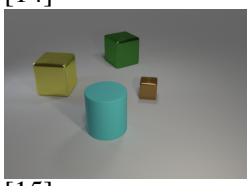
T$_A$: Hide all the cylinders from the scene.
Q$_H$: Are there any other things that have the same size as the gray sphere?
A: No
Classification: Remove action, Compare Attribute question
Split: val

[13]



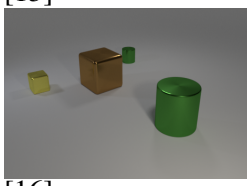
T$_A$: The small block is displaced and put on the left of the blue cube.
Q$_H$: Is there anything else on the right of the cyan sphere that has the same color as the large metal cylinder?
A: No
Classification: Move (in-plane) action, Compare Attribute question
Split: val

[14]



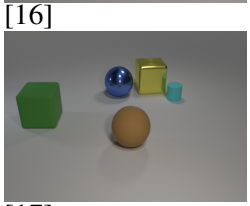
T$_A$: Jill places the small cube on the large cube that is to the left of cyan cylinder.
Q$_H$: There is an object below the brown cube; does it have the same shape as the green object?
A: Yes
Classification: Move (out-of-plane) action, Compare Attribute question
Split: val

[15]



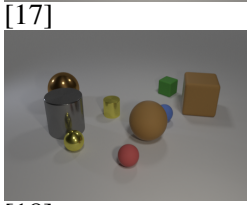
T$_A$: A small brown cube is added to the scene which is made of same material as the golden block.
Q$_H$: Are there an equal number of green objects and brown cubes?
A: Yes
Classification: Add action, Compare Integer question
Split: val

[16]



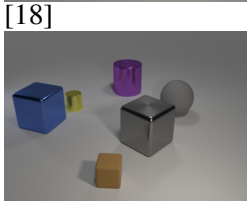
T$_A$: The tiny cylinder is withdrawn from the scene.
Q$_H$: Is the number of rubber objects greater than the number of shiny objects?
A: No
Classification: Remove action, Compare Integer question
Split: val

[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

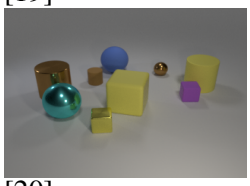
[18]



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

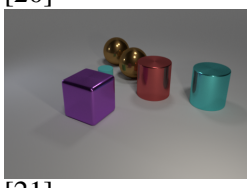
Figure 11: More examples from the CLEVR_HYP dataset

[19]



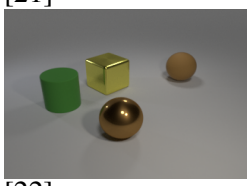
T$_A$: John hides the big object to the right of the brown sphere.
Q$_H$: How many yellow or cyan objects are there?
A: 3
Classification: Remove action, Counting question with 'Or'
Split: 2HopQ$_H$ test

[20]



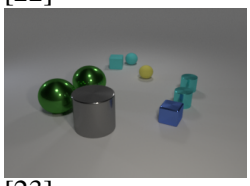
T$_A$: All brown things become matte.
Q$_H$: How many any other things are there which are made of the same material as the small cyan object?
A: 2
Classification: Change action, Counting + Compare Attribute question
Split: 2HopQ$_H$ test

[21]



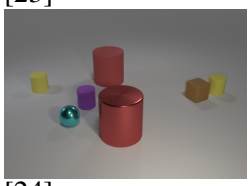
T$_A$: Make all the brown objects shiny.
Q$_H$: Are there any non metal things to the right of the shiny sphere?
A: No
Classification: Change action, Existence question with negation
Split: 2HopQ$_H$ test

[22]



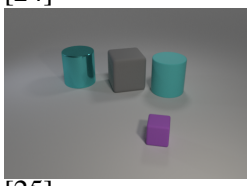
T$_A$: The gray object is moved to the right of the yellow thing.
Q$_H$: There is a cyan block; what number of big objects are there to the left of it that has the same material as the blue cube?
A: 2
Classification: Move (in-plane) action, Counting + Compare Attribute question
Split: 2HopQ$_H$ test

[23]



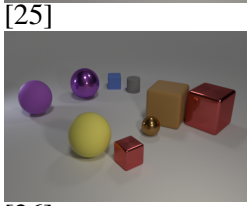
T$_A$: Remove all the yellow cylinders; Then clone the brown object and put it to the left side of the cyan ball.
Q$_H$: How many objects are made of the rubber?
A: 4
Classification: Add+Remove actions, Count question
Split: 2HopT$_A$ test

[24]



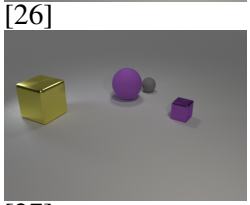
T$_A$: Enlarge the purple object; Then add a large red matte sphere to the right of the large purple cube.
Q$_H$: Is there any small object in the scene?
A: No
Classification: Add+Change actions, Existence question
Split: 2HopT$_A$ test

[25]



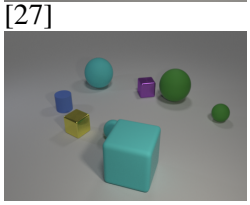
T$_A$: Add a small brown rubber sphere to the left of yellow matte object; Then swap its position with the purple shiny sphere.
Q$_H$: There is a ball that is to the left of the blue cube; what is its color?
A: Brown
Classification: Add+Move actions, Query Attribute question
Split: 2HopT$_A$ test

[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