

# Transformation Driven Visual Reasoning

Xin Hong<sup>1,2</sup> Yanyan Lan<sup>3,\*</sup> Liang Pang<sup>1,2</sup> Jiafeng Guo<sup>1,2</sup> Xueqi Cheng<sup>1,2</sup>

<sup>1</sup> CAS Key Laboratory of Network Data Science and Technology,

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

<sup>2</sup> University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup> Institute for AI Industry Research, Tsinghua University, Beijing, China

{hongxin19b, pangliang, guojiafeng, cxq}@ict.ac.cn lanyanyan@tsinghua.edu.cn

## Abstract

This paper defines a new visual reasoning paradigm by introducing an important factor, i.e. transformation. The motivation comes from the fact that most existing visual reasoning tasks, such as CLEVR in VQA, are solely defined to test how well the machine understands the concepts and relations within static settings, like one image. We argue that this kind of **state driven visual reasoning** approach has limitations in reflecting whether the machine has the ability to infer the dynamics between different states, which has been shown as important as state-level reasoning for human cognition in Piaget’s theory. To tackle this problem, we propose a novel **transformation driven visual reasoning** task. Given both the initial and final states, the target is to infer the corresponding single-step or multi-step transformation, represented as a triplet (object, attribute, value) or a sequence of triplets, respectively. Following this definition, a new dataset namely TRANCE is constructed on the basis of CLEVR, including three levels of settings, i.e. Basic (single-step transformation), Event (multi-step transformation), and View (multi-step transformation with variant views). Experimental results show that the state-of-the-art visual reasoning models perform well on Basic, but are still far from human-level intelligence on Event and View. We believe the proposed new paradigm will boost the development of machine visual reasoning. More advanced methods and real data need to be investigated in this direction. The resource of TVR is available at <https://hongxin2019.github.io/TVR>.

## 1. Introduction

Visual reasoning is the process of solving problems on the basis of analyzing the visual information, which goes well beyond object recognition [11, 21, 32, 33]. Though

\*Corresponding author.

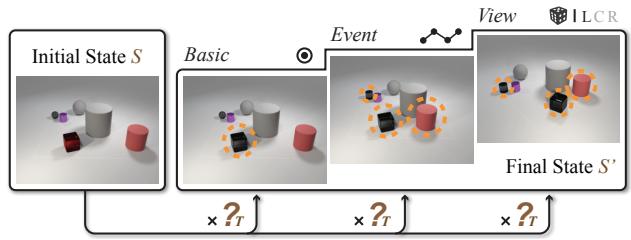


Figure 1. Illustration of three settings in TRANCE. **Basic:** Find the single-step transformation between the initial and final state. **Event:** Find the multi-step transformation between two states. **View:** Similar to Event but the view angle of the final state is randomly chosen from *Left*, *Center* (Default), and *Right*.

the task is easy for the human, it is tremendously difficult for vision systems, because it usually requires higher-order cognition and reasoning about the world. In recent years, several visual reasoning tasks have been proposed and attract a lot of attention in the community of computer vision, machine learning, and artificial intelligence. For example, the most representative visual question answering (VQA) tasks, such as CLEVR [18], define a question answering paradigm to test whether machines have spatial, relational, and other reasoning abilities for a given image. Visual entailment tasks such as NLVR [35, 36] ask models to determine whether a sentence is true about the states of two images. Visual commonsense reasoning tasks, such as VCR [45], further require the model to provide a rationale explaining why its answer is right.

We can see that these visual reasoning tasks are all defined at *state* level. For example, the language descriptions in NLVR as well as the questions and answers in VQA and VCR are just related to the concepts or relations within states, i.e. an image, or two images. We argue that this kind of **state driven visual reasoning** fails to test the ability of reasoning dynamics between different states. Take two images as an example. In the first image, there is a cat on a

tree, and in the second image, the same cat is under the tree. It is natural for a human to reason that the cat jumps down the tree after analyzing the two images. Piaget’s cognitive development theory [31] describes the dynamics between states as transformation, and tells that human intelligence must have functions to represent both the transformational and static aspects of reality. In addition, transformation is the key to tackle some more complicated tasks such as storytelling [13] and visual commonsense inference [29]. Though these tasks are closer to reality, they are too complicated to serve as a good testbed for transformation based reasoning. Because many other factors like representation and recognition accuracy may have some effects on the performance. Therefore, it is crucial to define a specific task to test the transformation reasoning ability.

In this paper, we define a novel *transformation driven visual reasoning* (TVR) task. Given the initial and final states, like two images, the goal is to infer the corresponding single-step or multi-step transformation. Without loss of generality, in this paper, transformations indicate changes of object attributes, so a single-step and multi-step transformation are represented as a triplet (*object, attribute, value*) and a sequence of triplets, respectively.

Following the definition of TVR, we construct a new dataset called TRANCE (Transformation on CLEVR), to test and analyze how well machines can understand the transformation. TRANCE is a synthetic dataset based on CLEVR [18], since it is better to first study TVR in a simple setting and then move to more complex real scenarios, just like people first study VQA on CLEVR and then generalize to more complicated settings like GQA. CLEVR has defined five types of attributes, i.e. color, shape, size, material, and position. Therefore, it is convenient to define the transformation for each attribute, e.g. the color of an object is changed from red to blue. Given the initial and final states, i.e. two images, where the final state is obtained by applying a single-step or multi-step transformation on the initial state, a learner is required to well infer such transformation. To facilitate the test for different reasoning levels, we design three settings, i.e. Basic, Event, and View. Basic is designed for testing single-step transformation. Event and View are designed for more complicated multi-step transformation reasoning, where the difference is that View further considers variant views in the final state. Figure 1 gives an example of three different settings.

In the experiments, we would like to test how well existing reasoning techniques [14, 19] work on this new task. However, since these models are mainly designed for existing reasoning tasks, they cannot be directly applied to TRANCE. To tackle this problem, we propose a new encoder-decoder framework named TranceNet, specifically for TVR. With TranceNet, existing techniques can be conveniently adapted to TVR. We test several differ-

ent encoders, e.g. ResNet [12], Bilinear-CNN [23] and DUDA [28]. While for the decoder, an adapted GRU [6] network is used to employ the image features and additional object attributes from TRANCE to predict the transformation, which is a sequence of triplets. Experimental results show that deep models perform well on Basic, but are far from human’s level on Event and View, demonstrating high research potentials in this direction.

In summary, the contributions of our work include: 1) the definition of a new visual reasoning paradigm, to learn the dynamics between different states, i.e. transformation; 2) the proposal of a new dataset called TRANCE, to test three levels of transformation reasoning, i.e. Basic, Event, and View; 3) experimental studies of the existing SOTA reasoning techniques on TRANCE show the challenges of the TVR and some insights for future model designs.

## 2. Related Works

The most popular visual reasoning task is VQA. Questions in the earliest VQA dataset [2, 10, 46] are usually concerned about the category or attribute of objects. Recent VQA datasets have improved the requirements on image understanding by asking more complex questions, e.g. Visual7W [47], CLEVR [18], OK-VQA [26], and GQA [15]. There are two other forms of visual reasoning tasks that need to be mentioned. Visual entailment tasks, such as NLVR [35, 36], and SNLI-VE [42, 43], ask models to determine whether a given description is true about a visual input. Visual commonsense reasoning [38, 45] tasks require to use commonsense knowledge [37] to answer questions. It is meaningful to solve these tasks which require various reasoning abilities. However, all the above tasks are defined to reason within a single state, which ignore the dynamics between different states.

Recently, several visual reasoning tasks have been proposed to consider more than one state. For example, CATER [9] tests the ability to recognize compositions of object movements. While our target is to evaluate transformations, and our data contains more diverse transformations rather than just moving. Furthermore, CATER along with other video reasoning tasks such as CLEVRER [44] and physical reasoning tasks [3, 4] are usually based on dense states, which make the transformations hard to define and evaluate. Before moving to these complex scenarios, our TVR provides a simpler formulation by explicitly defining the transformations between two states, which is more suitable for testing the ability of transformation reasoning. CLEVR-Change [28] is the most relevant work, which requires to caption the change between two images. The novelty is that TVR isolates the ability to reason about state-transition dynamics and supports a more thorough evaluation than captioning. Furthermore, CLEVR-Change only focuses on single-step transformations.

The concept of transformation has also been mentioned in many other fields. In [16, 22, 27], transformations are used to learn good attribute representations to improve the classification accuracy. In [1, 8, 25, 39, 48], object or environment transformations are detected to improve the performance of action recognition. However, those works in attribute learning and action recognition fields only consider single step transformation, thus not appropriate for testing a complete transformation reasoning ability. Some people may feel that procedure planning [5] has a similar task formulation to TVR and they are closer to reality. However, actions in planning data are usually very sparse. More importantly, both high-quality recognition and transformation reasoning are crucial to well model the task. As they are coupled together, it is not appropriate to use such tasks to evaluate the ability of transformation reasoning. That is why in this paper we use a more simple yet effective way to define the TVR task.

### 3. The Definition of TVR

TVR (Transformation driven Visual Reasoning) is a visual reasoning task that aims at testing the ability to reason the dynamics between states. Formally, we denote the state and transformation space as  $\mathcal{S}$  and  $\mathcal{T}$  respectively. The process of transforming the initial state into the final state can be illustrated as a function  $f : \mathcal{S} \times \mathcal{T} \rightarrow \mathcal{S}$ . Therefore, the task of TVR can be defined as inferring the transformation  $T \in \mathcal{T}$  given both the initial state  $S \in \mathcal{S}$  and final state  $S' \in \mathcal{S}$ . The space of transformation is usually very large, e.g. any changes of pixel value can be treated as a transformation. Therefore, without loss of generality, we define the atomic transformation as an attribute-level change of an object, represented as a triplet  $t = (o, a, v)$ , which means the object  $o$  with the attribute  $a$  is changed to the value  $v$ . For example, the color of an object is changed to blue. And further taking the order of atomic transformations into account, the transformation can be formalized as a sequence of atomic transformations, denotes as  $T = \{t_1, t_2, \dots, t_n\}$ , where  $n$  is the number of atomic transformations.

We make a distinction between the *single-step* ( $n = 1$ ) and *multi-step* ( $n \geq 1$ ) transformation setting because they can be evaluated in different ways to reflect different levels of transformation reasoning abilities. For single-step problems, we can directly compare the prediction  $\hat{T}$  with the ground-truth  $T$  to obtain overall accuracy, as well as the fine-grained accuracy of each element in the triplet  $(o, a, v)$ . With these metrics, it is convenient to know how well a learner understands atomic transformations and to analyze why a learner performs not well. However, multi-step transformation problems cannot be evaluated in this way. This is because atomic transformations sometimes are independent so that the order of them can be varying, which makes the answer not unique anymore. For example, the procedures

Size $\times 3$ Shape $\times 3$ Material $\times 3$ Color $\times 8$ Position $\times 16$ distance $\times 2$	S small* M medium L large* cube* sphere* cylinder* glass metal* rubber* yellow* gray* cyan* blue* brown* green* red* purple* front behind left right front right behind left front left behind 1 step 2 step	Total Attributes: 5 Total Values: 33 *: existing values in CLEVR
---	---	--

Table 1. Attributes and values in TRANCE.

of cooking can not be disrupted but changing the order of cooking different dishes makes no difference to the final result. To tackle this multi-solution problem, we consider the reconstruction error as the evaluation metric. Specifically, the predicting transformation  $\hat{T}$  is first applied to the initial state  $S$  to obtain the predicted final state  $\hat{S}'$ . Then the transformed  $\hat{S}'$  is compared with the ground-truth final state  $S'$  to decide whether the predicting transformation is correct and how far is the prediction from a correct transformation. Therefore, the evaluation of multi-step problems focuses more on the ability to find all atomic transformations and a feasible order to arrange them.

With this definition, most existing state driven visual reasoning tasks can be extended to the corresponding transformation driven ones. For example, the VQA task, such as CLEVR, can be extended to ask the transformation between two given images, with answers as the required transformation. In the extension of NLVR, the task becomes to determine whether a sentence describing the transformation is true about the two images, e.g. the color of the bus is changed to red. Since TVR itself is defined as an interpretation task, we do not need any further rational explanations, and the extension of VCR will stay the same as NLVR. We can see that the intrinsic reasoning target of these tasks is the same, that is to infer the correct transformation. While the difference lies in the manifestation.

### 4. The TRANCE Dataset

In this paper, we extend CLEVR by asking a uniform question, i.e. what is the transformation between two given images, to test the ability of transformation reasoning. This section introduces how the TRANCE (Transformation on CLEVR) dataset built following the definition of TVR.

#### 4.1. Dataset Setups

We choose CLEVR [18] to extend because CLEVR defines multiple object attributes, which can be changed conveniently. With the powerful Blender [7] engine used by CLEVR, we are able to collect over 0.5 million samples with only computational costs.

According to our definition, an atomic transformation is

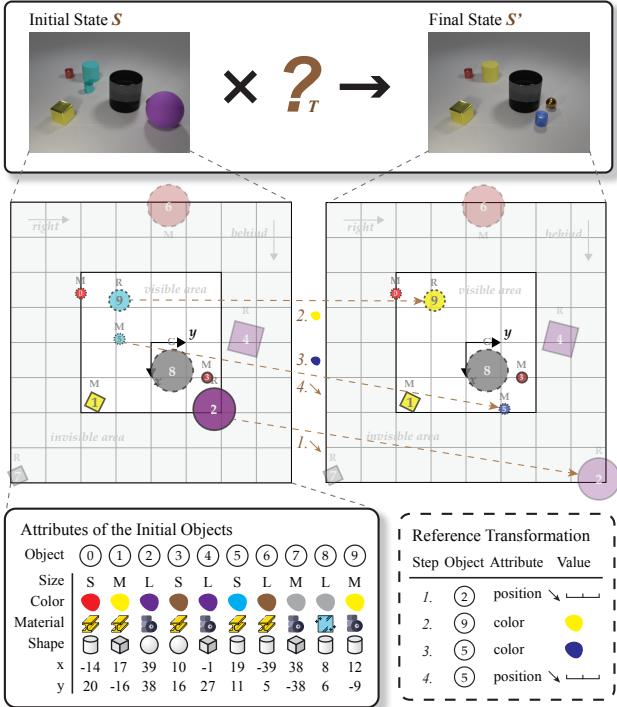


Figure 2. An example from the Event setting.

represented as a triplet  $(o, a, v)$ . In the following, we introduce attribute, value, and object one by one to demonstrate how to ground these factors.

The setup of the attribute is exactly the same as CLEVR. There are five attributes for each object, i.e. size, color, shape, material, and position.

The value in an atomic transformation is defined corresponding to the concerned attribute, as shown in Table 1. The values of the attributes except for position are similar to the default setting of CLEVR. Medium size and glass material are added to enrich the values. The value of position is carefully designed since it can be infinite in the space  $\mathbb{R}^2$ . To reduce the computation, we replace the absolute value of the position with a relative one by using direction and step to represent the value of a position transformation. We consider eight values for the variable direction, as shown in the Table 1. Then we define a coordinate system, in which  $x$  and  $y$  are both restricted to  $[-40, 40]$ , and objects can only be placed on integer coordinates. The variable step can be valued as 1 or 2, where 1 step equals 10 in our coordinate system. Except for normal moving action, we are also interested in whether the vision system could understand actions like moving in and moving out, so the plane is split into the visible area and the invisible area as shown in the middle two images of Figure 2, and the moving in and out operations can be defined correspondingly. To be reasonable, objects shouldn't be overlapped and moved out of the plane during transformation.

The setup of the object is basically the same as CLEVR. The only problem is how to represent an object in the answer. Existing methods such as CLEVR and CLEVR-Change use text which has ambiguity issues making the evaluation unreliable, while CLEVR-Ref+ [24] employs bounding boxes which is specific but requires the additional ability of detection. Therefore, we propose a simple method that is specific and easy to evaluate by providing the attributes of the initial objects, as shown in Figure 2. In this way, an object can be referred to with the assigned number. Note machines still need to perform their own recognition to align objects in images with given attributes.

To generate TRANCE, the first step is the same as CLEVR, which is randomly sampling a scene graph. According to the scene graph, CLEVR then generates questions and answers with a functional program and renders the image with Blender. Different from CLEVR, the next step in TRANCE generation becomes randomly sampling a sequence of atomic transformations ( $n \in \{1, 2, 3, 4\}$ ), which is called as the reference transformation, to transform the initial scene graph to the final scene graph. At last, two scene graphs are rendered into images ( $h : 240 \times w : 320$ ). The attributes of the initial objects can be easily obtained from the initial scene graph.

To reduce the potential bias in TRANCE, we carefully control the sampling process of scene graph and transformation by balancing several factors. In scene graph sampling, we balance objects' attributes and the number of visual objects in the initial state. In transformation sampling, the length of the transformation, the object number, n-gram atomic transformation, and the move type are all balanced. Throughout all elements, N-gram atomic transformation is the hardest to be balanced and it refers to the sub-sequence of atomic transformations with the length of  $n$ . By balancing these factors, we reduce the possibility that a learner utilizes statistics features in the data to predict answers. In the supplementary material, we show the statistics of the dataset and our balancing method in detail.

## 4.2. Three Levels of Settings

To facilitate the study on different levels of transformation reasoning, we design three settings, i.e. Basic, Event, and View. Basic is designed for single-step transformation and Event is for multi-step transformation. To further evaluate the ability of reasoning transformation under a more real condition, we extend Event with variant views to propose View. Figure 1 compares three different settings, more examples can be found in the supplementary material.

**Basic.** Basic is set as the first simple problem to mainly test how well a learner understands atomic transformations. The target of Basic is to infer the single-step transformation between the initial and final states. That is, given a pair of images, the task is to find out which attribute  $a$  of which

object  $o$  has been changed to which value  $v$ . We can see that this task is similar to the previous game ‘Spot the Difference’ [17], in which the player is asked to point out the differences between two images. However, Basic is substantially different from the game. The game focuses on the pixel level differences while Basic cares about the object level differences. Therefore, Basic can be viewed as a more advanced visual reasoning task than the game.

**Event.** It is obviously not enough to consider only the single-step transformation. In reality, it is very common that multi-step transformation exists between two states. Therefore, we construct this multi-step transformation setting to test whether machines can handle this situation. The number of transformations between the two states is randomly set from 1 to 4. The goal is to predict a sequence of atomic transformations that could reproduce the same final state from the initial state. To resolve this problem, a learner must find all atomic transformations and arrange them with a feasible order. Compared with Basic, it is possible to have multiple transformations, which improves the difficulty of finding them all. Meanwhile, the order is essential in the Event setting because atomic transformations may be dependent. For example, in Figure 2, two moving steps, i.e. 1st step and 4th step, cannot be exchanged, otherwise, object 5 and object 2 will overlap.

**View** The view angle of Basic and Event is fixed, which is not the case in real applications. To tackle this problem, we extend the Event setting to View, by capturing two states with cameras in different positions. In practice, for simplicity but without loss of generality, we set three cameras, placed in the left, center, and right side of the plane. The initial state is always captured by the center camera, while for the final state, images are captured with all three cameras. Thus, for each sample, we obtain three pairs for training, validation, and testing with the same initial state but different views of final states. In this way, we are capable to test whether a vision system can understand object-level transformation with variant views.

### 4.3. Evaluation Metrics

For the single-step transformation setting, i.e. Basic, the answer is unique. Therefore, we can evaluate the performance by directly comparing the prediction with the reference transformation, which is also the ground-truth transformation. Specifically in this paper, we consider two types of accuracy. The first one is fine-grained accuracy corresponds to three elements in transformation triplet, including object accuracy, attribute accuracy, and value accuracy, denoted as  $ObjAcc$ ,  $AttrAcc$ , and  $ValAcc$  respectively. The other one is the overall accuracy, which only counts the absolutely correct transformation triplets, denoted as  $Acc$ .

For multi-step transformation settings, i.e Event and View, it is not suitable to use the above evaluation met-

rics, because the answers may not be unique. Therefore, the predicting atomic transformation sequence is evaluated by checking whether it could reproduce the same final state as the reference transformation. Specifically, we first obtain the corresponding final state  $\hat{S}'$  by applying the predicting transformation  $\hat{T}$  to the initial state  $S$ , i.e.  $S \times \hat{T} \rightarrow \hat{S}'$ . Then a *distance* is computed by counting the attribute level difference between  $\hat{S}'$  and the ground-truth final state  $S'$ . To eliminate the influence of different transformation lengths on distance, we normalize it by the length of the reference transformation to get a *normalized distance*. Averaging these two metrics on all samples, we obtain  $AD$  and  $AND$ . We further consider the overall accuracy denoted as  $Acc$  when the distance equals to zero. In addition, we are interested to see without considering the order, whether all atomic transformations are found, by omitting all constraints such as no overlapping to compute the loose accuracy, which denotes as  $LAcc$ . At last, to measure the ability of assigning the right order when all atomic transformations have been found, the error of order  $EO = (LAcc - Acc)/LAcc$  is computed. In summary, five evaluation metrics are used in multi-step transformation settings, i.e.  $AD$ ,  $AND$ ,  $LAcc$ ,  $Acc$ , and  $EO$ .

In the evaluation of multi-step transformation problems, an important step is to obtain the predicting final state  $\hat{S}'$  by applying the predicting transformation  $\hat{T}$  to the initial state  $S$ . This function has already been implemented in the previously mentioned data generation system and we reuse it in our multi-step transformation evaluation system. Except for the usage of evaluation, this evaluation system can also be used to generate signals as rewards for reinforcement learning, which is explored in Section 5.2.

## 5. Experiments

In this section, we show our experimental results on the three settings of TRANCE, i.e. Basic, Event, and View. We also conduct analyses to show some insights about machines’ ability of reasoning transformation.

### 5.1. Models

Firstly, we would like to test how well existing methods work on this new task. However, since the inputs and outputs of TVR are quite different from existing visual reasoning tasks, existing methods like [14, 19] cannot be directly applied. So we design a new encoder-decoder style framework named TranceNet. As Figure 3 shows, start from the input image pairs, an encoder first extracts features, and then a GRU based decoder is employed to generate transformation sequences. The following of this section briefly introduces our TranceNet framework while the implementation details can be found in the supplementary material. To compare with the human, for each of the three settings, we also collect results of 100 samples in total. These results

come from 10 CS Ph.D. candidates who are familiar with our problems and the testing system.

**Encoder.** The goal of an encoder is to extract effective features from image pairs, which are mainly associated with the difference between the two states. There are two ways to extract these features, namely single-stream way and two-stream way. Single-stream way directly inputs two images into a network to extract features, while two-stream way first separately extracts image features and then interacts them in feature-level. In this paper, we evaluate six methods. In terms of the single-stream way, we test two networks, i.e. Vanilla CNN and ResNet [12], combined with two preprocessing strategies, i.e. subtraction ( $-$ ) and concatenation ( $\oplus$ ). For example, we use ResNet $_{\oplus}$  to represent a ResNet fed with concatenated image pairs. Another two methods are Bilinear CNN (BCNN [23]) and a recently proposed method called DUDA [28], which operate as the two-stream way. BCNN is a classical model for fine-grained image classification to distinguish categories with small visual differences. DUDA is originally proposed for change detection and captioning. The main difference between BCNN and DUDA lies in the way of feature-level interaction.

**Decoder.** The decoder is used to output a feasible transformation sequence from the extracted image features. All six encoders share the same modified GRU [6] network, which is a commonly used technique for sequence generation. As shown in Figure 3, the major difference between our GRU network and a standard one lies in the additional classifiers. A classifier unit accepts attributes of the initial objects and the current hidden state from the GRU cell, and then outputs the object and value of the current step. In detail, at a certain step, an object vector is first computed from the hidden state. Then the object number is obtained by matching the object vector with the initial objects using cosine similarity. Finally, the most similar object vector from initial objects and the hidden state are used to predict the value. In TRANCE, attributes are implied by values, for example, blue indicates that the attribute is color, so that the output of a classifier does not explicitly include an attribute.

Since all these models share the same decoder, we denote these models by their encoders' names hereafter.

**Training.** The loss function of a single sample consists of two cross-entropy losses for object and value respectively, which can be represented as:

$$\mathcal{L}(T, \hat{T}) = -\frac{1}{n} \sum_{i=1}^n (t_i^o \cdot \log \hat{t}_i^o + t_i^v \cdot \log \hat{t}_i^v), \quad (1)$$

where  $n$  is the transformation length of the sample,  $t_i^o$  and  $t_i^v$  denote the object and value in the  $i$ -th step of transformation. The training loss is the average of losses over all training examples. During training, we use teacher forcing [41] for faster convergence in two parts of TranceNet. Firstly, at each step, we follow the practice in sequence learning such

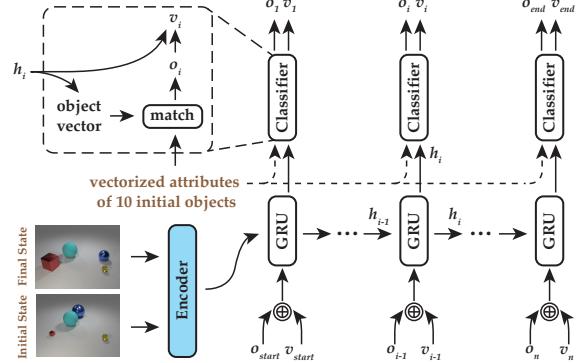


Figure 3. The architecture of TranceNet.

as machine translation by using the object and value from the given reference transformation as the inputs of the GRU unit. Additionally, in the classifier, objects from reference transformations are used to predict values.

## 5.2. Results on Three Settings

In this section, we first present our experimental results on the three settings of TRANCE, and then provide some in-depth analyses on the results.

Six models are tested in the Basic setting of TRANCE, i.e. CNN $_-$ , CNN $_{\oplus}$ , ResNet $_-$ , ResNet $_{\oplus}$ , BCNN, DUDA. From the results in the left part of Table 2, we can see that all models perform quite well, in the sense that the performance gap between these models and the human is not very large. Comparing these models, both versions of ResNet perform better than BCNN and DUDA. As we mentioned before, CNN and ResNet are single-stream methods while BCNN and DUDA are two-stream methods. Since the model size of ResNet, BCNN, and DUDA is similar, we can conclude that the single-stream way is better than the two-stream way on the Basic setting. Further checking the fine-grained accuracy, we can see this gap comes from the ability to find the correct objects and values, while all models are good at distinguishing different attributes.

The experimental results on Event are shown in the middle part of Table 2. We can see that this task is very challenging for machines, since there is an extremely big performance gap between models and the human. That is because the answer space rises exponentially when the number of steps increases. In our experiments, the size of answer space is  $\sum_{i=1}^4 (33 \times 10)^i$ , about 11.86 billion. The performance (e.g. Acc) gap between CNN and ResNet models becomes larger from Basic to Event, which suggests larger encoders have advantages in extracting sufficient features to decode transformation sequences.

We also employ reinforcement learning to train models. Specifically, the evaluation system introduced in Section 4.3 can provide signals include the *correctness* of a prediction and the *distance* of a prediction to the correct

Model	Basic				Event				View			
	ObjAcc↑	AttrAcc↑	ValAcc↑	Acc↑	AD↓	AND↓	LAcc↑	Acc↑	AD↓	AND↓	LAcc↑	Acc↑
CNN_	0.9584	0.9872	0.9666	0.9380	1.5475	0.5070	0.4587	0.4442	2.2376	0.8711	0.2344	0.2286
CNN <sub>⊕</sub>	0.9581	0.9889	0.9725	0.9420	1.4201	0.4658	0.4981	0.4838	2.2517	0.8764	0.2350	0.2285
ResNet_	0.9830	0.9969	<b>0.9935</b>	0.9796	1.0974	<b>0.3417</b>	0.5972	0.5750	<b>1.1068</b>	0.3749	0.5484	0.5272
ResNet <sub>⊕</sub>	<b>0.9852</b>	<b>0.9980</b>	0.9928	<b>0.9810</b>	<b>1.0958</b>	0.3469	<b>0.6019</b>	<b>0.5785</b>	1.1148	<b>0.3731</b>	<b>0.5525</b>	<b>0.5305</b>
BCNN	0.9705	0.9950	0.9788	0.9571	1.1081	0.3582	0.5746	0.5560	1.2633	0.4395	0.4977	0.4784
DUDA	0.9453	0.9888	0.9692	0.9320	1.5261	0.4975	0.5025	0.4856	1.5352	0.5242	0.4746	0.4590
Human	1.0000	1.0000	1.0000	1.0000	0.3700	0.1200	0.8300	0.8300	0.3200	0.0986	0.8433	0.8433

Table 2. Model and human performance on Basic, Event, and View.

Model	AD↓	AND↓	LAcc↑	Acc↑
ResNet <sub>⊕</sub>	1.0958	0.3469	0.6019	0.5785
+ corr	1.0579	0.3316	0.6215	0.5978
+ dist	1.0528	0.3319	0.6180	0.5938
+ corr & dist	<b>1.0380</b>	<b>0.3251</b>	<b>0.6230</b>	<b>0.6001</b>

Table 3. Results of ResNet<sub>⊕</sub> trained using REINFORCE [40] with different rewards on Event.

one. These signals are able to be used as rewards in REINFORCE [40] algorithm to further train ResNet<sub>⊕</sub> models. Table 3 shows that all three rewards significantly improve the performance, and the difference among them is small.

The right part of Table 2 shows the results on the View setting. While humans are insensitive to view variations, the performances of all deep models drop sharply from Event to View. Among these models, the most robust one is DUDA. From the fact that BCNN has a similar two-stream architecture but performs worse, we can see that DUDA’s way of directly interacting two state features is more effective for tackling the view variation.

### 5.3. Detailed Analysis on Event and View

According to the above experimental results, the performances on Event and View are not good. So we conduct some detailed analysis to help understand the task and provide some insights for future model designs.

Firstly, we analyze the effect of transformation sequence length on Event, which is the main factor to make performance worse than Basic. Specifically, we separate all test samples into four groups based on their lengths, i.e. samples with  $k$ -step transformation ( $k = 1, 2, 3, 4$ ). Then we plot the LAcc for each group in Figure 4. From the results, we can see that both human and deep models work quite well when the length is short, e.g. 1. As the length increases, humans still have the ability to well capture the complicated transformations. However, the deep models decline sharply. Take CNN<sub>–</sub> as an example, the performances for the four different groups are 95%, 56%, 23%, and 10%. These re-

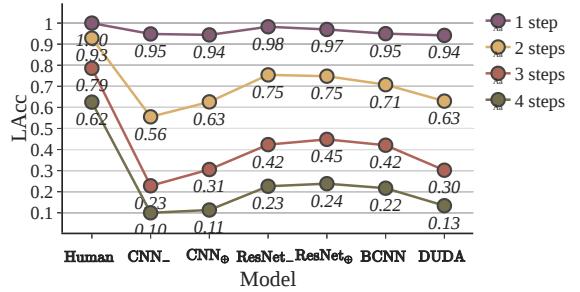


Figure 4. Results on Event with respect to different steps.

Encoder	LAcc↑	Acc↑	EO↓
Random (avg. of 100)	1.0000	0.4992	0.5008
CNN <sub>–</sub>	0.2276	0.2067	<b>0.0915</b>
CNN <sub>⊕</sub>	0.2596	0.2244	0.1358
ResNet <sub>–</sub>	0.3478	0.2997	0.1382
ResNet <sub>⊕</sub>	<b>0.3862</b>	<b>0.3205</b>	0.1701
BCNN	0.2949	0.2612	0.1141
DUDA	0.2500	0.2147	0.1410
Human	0.7273	0.7273	0.0000

Table 4. Results on 7.8% order sensitive samples from Event.

sults indicate that future studies should focus more on how to tackle transformations with long steps.

Then we analyze the effect of the order on Event, which is another important factor in this data. According to our statistics, there are about 7.8% test samples<sup>1</sup> exist certain permutations that violate our constraints such as no overlapping. That is to say, these samples are order sensitive. Even if a model is able to find all correct atomic transformations, the result still could be wrong without carefully considering an order to arrange them. Table 4 shows the results on these order sensitive samples, where EO is directly defined to measure the influence of order, LAcc and Acc are just listed for reference. From the results, we can see that

<sup>1</sup>We have also tried other data with different percentages, e.g. 25%, and the results are similar.

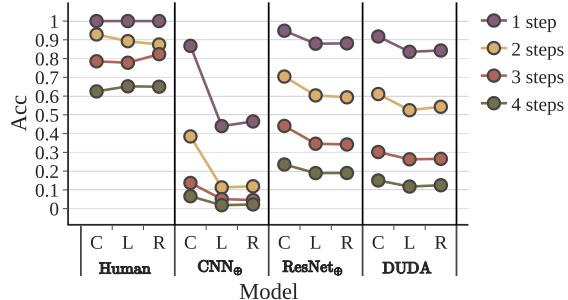


Figure 5. Results for different final views (Center, Left, Right).

EO of the human is zero. Therefore, once humans find all correct atomic transformations, it is easy to figure out the orders meanwhile. However, for all deep models, the EOs are larger than zero, which indicates a clear effect of the order on the reasoning process. In order to find out the extent of the effect, i.e., whether  $0.0915 \sim 0.1701$  means a large deviation, we perform an experiment on 100 randomly selected order sensitive samples. Specifically for each sample, we randomly assign an order for ground truth atomic transformations. As a result, the EO is 0.5008, which could be viewed as an upper bound of the order error. Therefore, the current deep models indeed have some ability to tackle the orders, but there still leaves some room for improvements.

At last, we analyze the effect of view variation. For each model, we provide the results of different final views, as shown in Figure 5. Please note that the results of CNN<sub>⊖</sub>, ResNet<sub>⊖</sub>, and BCNN are quite similar to CNN<sub>⊕</sub>, ResNet<sub>⊕</sub> and DUDA, so we just give the results from latter three typical models. Firstly, the results of humans across different views change small, demonstrating human’s powerful ability of adapting to different views. In some cases, humans perform even better when views are changed than unchanged. That is because when the view is altered, humans usually spend more time solving the problems, which decreases the chances of making errors. Conversely, deep learning models share a similar trend that view variations will hurt the performance. Among these models, CNN decreases the most, while DUDA shows its robustness. In conclusion, models with more parameters are more robust to view variations and feature-based interaction like the way used in DUDA, is helpful.

#### 5.4. Analysis of Training Data Size

In our experiments, we find that data size is an important factor for training and evaluation. Therefore, we use ResNet<sub>⊕</sub> as an example to study the influence of this factor. From Figure 6, we can see that more training samples bring significant benefits when the number is less than 50k on Basic and 200k on Event and View. After that, the benefits become smaller and smaller. Those results are consistent with the common knowledge that relatively large data

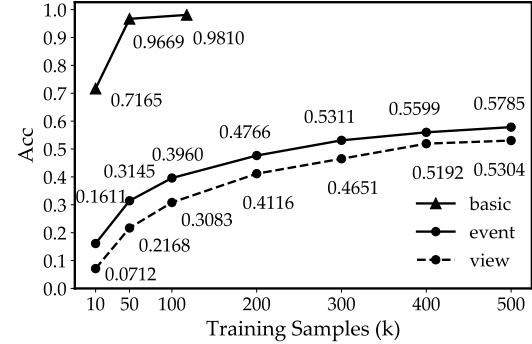


Figure 6. Results of ResNet<sub>⊕</sub> with different training size.

is required to well train a deep model. These results also show TRANCE has sufficient size in our experiments.

## 6. Conclusion

To tackle the problem that most existing visual reasoning tasks are solely defined on static settings and cannot well capture the dynamics between states, we propose a new visual reasoning paradigm, namely transformation driven visual reasoning (TVR). Given the initial and final states, the target is to infer the corresponding single-step transformation or multi-step transformations, represented by a triplet (object, attribute, value) or a sequence of triplets, respectively. In this paper, as an example, we use CLEVR to construct a new synthetic data, namely TRANCE, which includes three different levels of settings, i.e. Basic for single-step transformation, Event for multi-step transformation, and View for multi-step transformation with variant views. To study the effectiveness of existing SOTA reasoning techniques, we propose a new encoder-decoder framework named TranceNet and test six models under this framework. The experimental results show that our best model works well on Basic, while still has difficulties solving Event and View. Specifically, the difficult point of Event is to find all atomic transformations and arrange them with a feasible order, especially when the length of the sequence is large. While for View, the view variations bring great challenges to these models, but have little impact on humans.

In the future, we plan to investigate from both model and data perspectives by testing methods like neural symbolic approaches and constructing a real dataset to study TVR.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61773362, and 61906180, Beijing Academy of Artificial Intelligence (BAAI) under Grants BAAI2020ZJ0303, the National Key R&D Program of China under Grants No. 2016QY02D0405, the Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202033).

## Supplementary Material

This document aims at providing additional materials to supplement our main submission. We first show the detail of data balancing on TRANCE in Section A. Then, we give more details on the implementation of the baseline models and training in Section B. Next in Section C, we describe the test system we used for collecting results from humans. Finally in Section D, we provide extra examples of three settings from TRANCE, i.e. Basic, Event, and View.

### A. Dataset Balance

Data balancing is an important factor to be considered when constructing TRANCE. Several factors are balanced in TRANCE, so that a learner is expected to reason the transformation without utilizing the biased features such as the length of transformation in data. Without considering the image rendering, the data generation process consists of two stages, i.e. sampling an initial scene graph and sampling a transformation sequence to transform the initial scene graph into the final scene graph. In the following of this section, we first introduce the factors that are balanced in these two stages and then describe the method we use.

When sampling the initial scene graph, the attributes of all objects and the number of visible objects are balanced. Recall that the plane is separated into the visible area and invisible area and only objects in the visible area appear in the image of the initial state. The two diagrams on the top row of Figure 7 show the statistics of these two factors. We can see that they are strictly balanced.

When sampling the transformation sequence, we balance four factors in total. The first factor is the length of transformation so that samples with different transformation lengths are equal in terms of size. The statistic result of the transformation length can be found in Figure 7 on the left side of the second row. The other three factors are all about the elements of atomic transformations including the object number for object and the move type and  $n$ -gram atomic transformation for the value. The object number is directly balanced over all samples and the result is shown in the middle of the second row in Figure 7. The move type is also strictly balanced and the statistics is shown in the right of the second row in Figure 7. As for the  $n$ -gram atomic transformation, it should be handled carefully, since for a specific initial scene graph, the availability of different atomic transformations is different. For example, changing the color of one object can always be successful, but changing the position of an object with a specific direction and step may be failed because of overlapping. Thus, the concurrence of different atomic transformations has different probability. For example, four atomic transformations on position will be less possible than four atomic transformations on color exist in one sequence. Therefore, we need to consider to balance the

	1-gram	2-gram	3-gram	4-gram
options	33	1,089	35,937	1,185,921
min	38,635	697	7	0
max	38,638	708	15	3
median	38,636	703	11	0
mean	38,636	702.5	10.64	0.1075
std	0.7714	2.2854	0.7880	0.3150

Table 5. The statistics of the  $n$ -gram atomic transformations in TRANCE.

value throughout the sub-sequence of atomic transformations. In the following, we will call a sub-sequence with the length  $n$  as a  $n$ -gram atomic transformation. Table 5 shows the statistics of this factor. For each  $n$ -gram, the number of different options to be chosen is shown in the first row. For example, we have 33 different values so that the options of 1-gram are 33 and that of 2-gram is  $33^2 = 1089$  and so on. The process of counting  $n$ -gram atomic transformations can be regarded as counting the sub-sequences using a  $n$ -length sliding window with 1 stride on transformation sequences. For example, to count 2-gram atomic transformations on a 4-step transformation, we use a 2-length sliding window with 1 stride and there will be three 2-gram atomic transformations. The rows below the options in Table 5 are calculated on the counting results. From the table, the standard variance is very small compared to the mean value, which means the concurrence of different atomic transformations is well balanced. It should be note that the size of TRANCE is 0.5 million, which is not enough to cover all 4-gram options, but the analysis of training data size in our experiments has proved our data is sufficient for training a deep model. In conclusion, the dataset is well balanced to eliminate the potential bias that are not related to the target of transformation reasoning.

The method we used to balance all the above factors is called balanced sampling. The basic idea of this method is changing the probability of the sampling targets dynamically according to previously generated samples. Algorithm 1 shows how to sample an option given the count table of previously generated options.

### B. Implementation Details

The code for data generation is rewritten on the basis of the CLEVR<sup>2</sup>. In terms of training, we use PyTorch [30] as our deep learning framework. All of the code can be found at our Github repository<sup>3</sup>. In the following, we introduce the implementation of our baseline models and the training process in detail.

<sup>2</sup><https://github.com/facebookresearch/clevr-dataset-gen>

<sup>3</sup><https://github.com/hughplay/TVR>

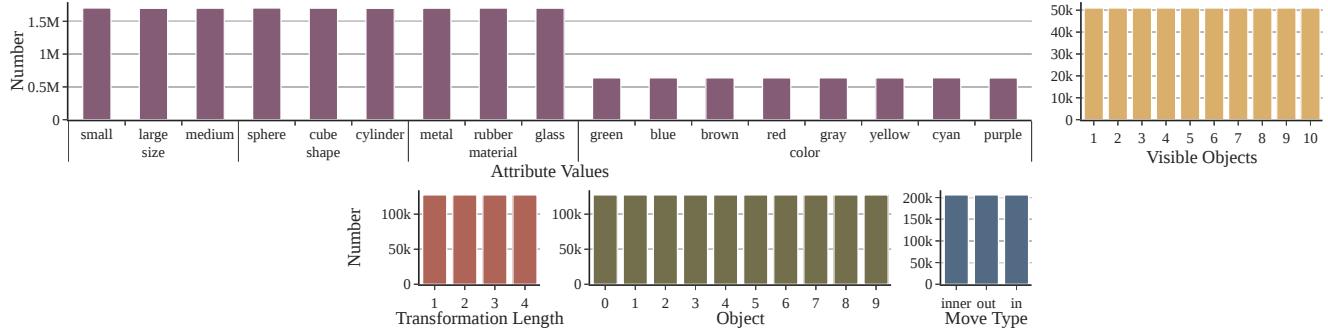


Figure 7. The statistics of balanced factors in the TRANCE dataset. **Top Row:** The attribute values and the visible objects which are balanced during sampling the initial scene graphs. **Bottom Row:** The transformation length, object number, and move type which are balanced when sampling transformation sequences.

---

**Algorithm 1:** Balanced Sampling

---

**Input:** available  $k$  options  $O = \{o_1, o_2, \dots, o_k\}$ ,  
corresponding count table  
 $N = \{n_1, n_2, \dots, n_k\}$ ;  
**Output:** sampled option  $o_r$ ;  
**Parameter:** tolerance  $t = 0.1$  (default);

- 1  $n_{max} = \max(n_1, n_2, \dots, n_k)$ ;
- 2  $c_i = n_{max} - n_i + t$ ;
- 3  $p_i = \frac{c_i}{\sum_{i=1}^k c_i}$ ;
- 4  $o_r$  = randomly sample an option from  $\{o_1, o_2, \dots, o_k\}$  with probability  $\{p_1, p_2, \dots, p_k\}$ ;

---

Table 6 shows the architectural details of different baseline models under the TranceNet framework. In the encoder part, both CNN<sub>-</sub> and CNN<sub>⊕</sub> use a 4-layer CNN as the backbone. The channel of four CNN layers is 16, 32, 32, 64, the kernel size is 5, 3, 3, 3, and all the strides is 2. The encoder backbone of ResNet<sub>-</sub>, ResNet<sub>⊕</sub>, and DUDA is ResNet-18 [12], which we directly use the implementation given by PyTorch without pretrained parameters. As for BCNN, we use the VGG-18 [34] implemented by PyTorch as the backbone, which is to keep consistent with the original paper [23]. In the decoder part, the output of the encoder is first flattened and then encoded by a fully-connected layer to become a 128-dimension vector. This 128-dimension vector will be sent to the adapted GRU network. In the GRU network, the hidden size of each GRU cell is 128 and two 1-layer fully-connected layers are used to decode the object vector and the value of each step respectively. The dimension of the object vector is 19 including 8 for the color, 3 for the size, 3 for the shape, 3 for the material, and 2 for the position. The dimension of the value output is 33.

The optimizer we used is Adam [20] and the learning rate is 0.001 in the beginning then reduced to 0.0001 after 25 epochs. The samples for Event and View is shared and the number of samples in the training, validation, and test set is

Model	Encoder Backbone	Decoder	Parameters
CNN <sub>-</sub>	4-layer CNN	Adapted GRU	737K
CNN <sub>⊕</sub>	4-layer CNN	Adapted GRU	738K
ResNet <sub>-</sub>	resnet18	Adapted GRU	11M
ResNet <sub>⊕</sub>	resnet18	Adapted GRU	11M
BCNN	vgg11.bn	Adapted GRU	41M
DUDA	resnet18	Adapted GRU	18M

Table 6. The architectural details of different baseline models under the TranceNet framework.

500,000, 10,000, and 20,000 respectively. Please note for each sample in View, we have 3 different final views so that the total image pairs are tripled. As for Basic, we collect all existing 1-step samples in data, and the size of training, validation, and test set is 125000, 2,500, and 5,000. All models are trained with 50 epochs on the training set and models that have the best results on the validation set are chosen to be evaluated on the test set to get the final results. In our experiments, images are resized to  $120 \times 160$  for fast training. Furthermore, by following the common practice on image augmentation [21], we apply a  $0 \sim 5\%$  spatial translation to all input image pairs during training.

During evaluation, when moving an object from the visible area into the invisible area, any directions and steps that could cause the same effect without making objects overlapping are accepted. In the evaluation system, this is implemented by only comparing the visible objects' attribute values of the two final states, i.e. the predicting final state and the ground truth final state.

## C. Human Test System

To collect the human results, we build a web-based test system. Figure 8 shows the GUI of this system. The whole testing process includes the following steps. First of all, a human tester is told to be familiar with our system by trying a few examples with guidance. After that, the tester can start

to test. During testing, for each sample, the tester should first observe given images and the attributes of the initial objects and then select the correct atomic transformations arranged with a feasible order. To reduce the time usage, we also provide the visualization of the initial objects for testers. After completing all samples, the tester can see his or her test result by checking the testing history.

## D. More Examples from TRANCE

The remaining pages show extra examples from the three settings of TRANCE, i.e. Basic, Event, and View. In each sample, the initial state, the final state, and the attributes of the initial objects are provided. In the View setting, the view of the final state is randomly selected from Left and Right. To make readers easy to understand the given examples, for each example, an additional diagram is provided to visualize the attributes of the initial objects. At last, we show the reference transformation.

**TVR**

Dataset Human Test

### Setting

Fill in the user name, select a problem, and then start to test.

User Name  Problem

### Task Brief

Given the initial state, the final state, and the attributes of the initial objects. The answer should be a transformation that could complete the change of states, which is a sequence of triplets with the length range from 1 to 4, while each triplet is represented as (<object, attribute, value>).

Notice: Overlapping and moving out of the invisible area is not allowed throughout the whole transformation process.

Initial State      Final State

Obj	Size	Color	Mat	Shape	Position
0	small	red	glass	cylinder	(12, -4)
1	large	brown	metal	cube	(0, -4)
2	small	gray	metal	sphere	(-26, 14)
3	small	yellow	rubber	cube	(-10, 17)
4	large	gray	rubber	cube	(10, 20)
5	small	brown	rubber	cube	(37, 22)
6	large	green	metal	sphere	(-10, -13)
7	small	blue	glass	sphere	(-11, 1)
8	small	cyan	metal	sphere	(19, 14)
9	medium	yellow	rubber	cylinder	(-8, 8)

The Attributes of the Initial Objects

### Your Answer

Select the correct transformation here. You could add a new or remove an existing atomic transformation with the corresponding buttons if you need. You can also change the existing order by dragging the handle of each column.

Order	Object	Attribute	Value
0	0	position	front-left.1
1	6	position	front-left.2

### Test Result

User: test  
Problem: view  
Index: 281452\_detail  
View Final: Camera\_Left  
Time Usage: 33 s

Correctness: ✓  
Ground Truth:

0	0	position.front-left.1
1	6	position.left.2

Your Answer:

0	0	position.front-left.1
1	6	position.front-left.2

### Visualization of the Initial Objects

This diagram visualizes the attributes of the initial objects.

Figure 8. Human test system.

	Initial State (Top) Final State (Bottom)	Attributes of the Initial Objects	Visualization of the Initial Attributes	Reference Transformation																																																																											
1		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>yellow</td><td>metal</td><td>cylinder</td><td>5,-18</td></tr> <tr><td>1</td><td>large</td><td>gray</td><td>rubber</td><td>sphere</td><td>19,40</td></tr> <tr><td>2</td><td>medium</td><td>gray</td><td>glass</td><td>cylinder</td><td>2,39</td></tr> <tr><td>3</td><td>small</td><td>cyan</td><td>rubber</td><td>cube</td><td>-4,7</td></tr> <tr><td>4</td><td>small</td><td>blue</td><td>rubber</td><td>cylinder</td><td>12,2</td></tr> <tr><td>5</td><td>large</td><td>yellow</td><td>metal</td><td>cylinder</td><td>13,-27</td></tr> <tr><td>6</td><td>medium</td><td>purple</td><td>rubber</td><td>cube</td><td>-10,-2</td></tr> <tr><td>7</td><td>small</td><td>yellow</td><td>glass</td><td>sphere</td><td>-21,30</td></tr> <tr><td>8</td><td>medium</td><td>blue</td><td>rubber</td><td>sphere</td><td>30,34</td></tr> <tr><td>9</td><td>small</td><td>gray</td><td>glass</td><td>sphere</td><td>-14,18</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	yellow	metal	cylinder	5,-18	1	large	gray	rubber	sphere	19,40	2	medium	gray	glass	cylinder	2,39	3	small	cyan	rubber	cube	-4,7	4	small	blue	rubber	cylinder	12,2	5	large	yellow	metal	cylinder	13,-27	6	medium	purple	rubber	cube	-10,-2	7	small	yellow	glass	sphere	-21,30	8	medium	blue	rubber	sphere	30,34	9	small	gray	glass	sphere	-14,18		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>material</td></tr> <tr><td>0</td><td>0</td><td>rubber</td></tr> </tbody> </table>	obj	attribute	value	0	0	material	0	0	rubber
obj	size	color	material	shape	position																																																																										
0	medium	yellow	metal	cylinder	5,-18																																																																										
1	large	gray	rubber	sphere	19,40																																																																										
2	medium	gray	glass	cylinder	2,39																																																																										
3	small	cyan	rubber	cube	-4,7																																																																										
4	small	blue	rubber	cylinder	12,2																																																																										
5	large	yellow	metal	cylinder	13,-27																																																																										
6	medium	purple	rubber	cube	-10,-2																																																																										
7	small	yellow	glass	sphere	-21,30																																																																										
8	medium	blue	rubber	sphere	30,34																																																																										
9	small	gray	glass	sphere	-14,18																																																																										
obj	attribute	value																																																																													
0	0	material																																																																													
0	0	rubber																																																																													
2		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>brown</td><td>glass</td><td>cylinder</td><td>2,-11</td></tr> <tr><td>1</td><td>small</td><td>blue</td><td>metal</td><td>cylinder</td><td>17,25</td></tr> <tr><td>2</td><td>large</td><td>red</td><td>metal</td><td>cylinder</td><td>8,12</td></tr> <tr><td>3</td><td>medium</td><td>yellow</td><td>glass</td><td>sphere</td><td>14,-7</td></tr> <tr><td>4</td><td>medium</td><td>purple</td><td>glass</td><td>cylinder</td><td>-26,37</td></tr> <tr><td>5</td><td>medium</td><td>gray</td><td>rubber</td><td>cube</td><td>-6,-13</td></tr> <tr><td>6</td><td>medium</td><td>blue</td><td>rubber</td><td>sphere</td><td>15,2</td></tr> <tr><td>7</td><td>medium</td><td>brown</td><td>rubber</td><td>cube</td><td>-20,10</td></tr> <tr><td>8</td><td>small</td><td>red</td><td>rubber</td><td>cube</td><td>37,-31</td></tr> <tr><td>9</td><td>small</td><td>purple</td><td>metal</td><td>cube</td><td>25,-37</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	brown	glass	cylinder	2,-11	1	small	blue	metal	cylinder	17,25	2	large	red	metal	cylinder	8,12	3	medium	yellow	glass	sphere	14,-7	4	medium	purple	glass	cylinder	-26,37	5	medium	gray	rubber	cube	-6,-13	6	medium	blue	rubber	sphere	15,2	7	medium	brown	rubber	cube	-20,10	8	small	red	rubber	cube	37,-31	9	small	purple	metal	cube	25,-37		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>2</td><td>color</td></tr> <tr><td>0</td><td>2</td><td>purple</td></tr> </tbody> </table>	obj	attribute	value	0	2	color	0	2	purple
obj	size	color	material	shape	position																																																																										
0	small	brown	glass	cylinder	2,-11																																																																										
1	small	blue	metal	cylinder	17,25																																																																										
2	large	red	metal	cylinder	8,12																																																																										
3	medium	yellow	glass	sphere	14,-7																																																																										
4	medium	purple	glass	cylinder	-26,37																																																																										
5	medium	gray	rubber	cube	-6,-13																																																																										
6	medium	blue	rubber	sphere	15,2																																																																										
7	medium	brown	rubber	cube	-20,10																																																																										
8	small	red	rubber	cube	37,-31																																																																										
9	small	purple	metal	cube	25,-37																																																																										
obj	attribute	value																																																																													
0	2	color																																																																													
0	2	purple																																																																													
3		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>yellow</td><td>glass</td><td>sphere</td><td>-34,10</td></tr> <tr><td>1</td><td>large</td><td>purple</td><td>metal</td><td>cylinder</td><td>-9,-8</td></tr> <tr><td>2</td><td>small</td><td>red</td><td>rubber</td><td>sphere</td><td>11,29</td></tr> <tr><td>3</td><td>small</td><td>yellow</td><td>glass</td><td>cube</td><td>-30,-20</td></tr> <tr><td>4</td><td>large</td><td>blue</td><td>rubber</td><td>cube</td><td>-13,35</td></tr> <tr><td>5</td><td>medium</td><td>yellow</td><td>glass</td><td>cylinder</td><td>-20,-37</td></tr> <tr><td>6</td><td>medium</td><td>red</td><td>rubber</td><td>sphere</td><td>17,-14</td></tr> <tr><td>7</td><td>small</td><td>red</td><td>metal</td><td>cylinder</td><td>-11,-26</td></tr> <tr><td>8</td><td>medium</td><td>yellow</td><td>glass</td><td>cube</td><td>-23,21</td></tr> <tr><td>9</td><td>small</td><td>blue</td><td>metal</td><td>cube</td><td>40,-23</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	yellow	glass	sphere	-34,10	1	large	purple	metal	cylinder	-9,-8	2	small	red	rubber	sphere	11,29	3	small	yellow	glass	cube	-30,-20	4	large	blue	rubber	cube	-13,35	5	medium	yellow	glass	cylinder	-20,-37	6	medium	red	rubber	sphere	17,-14	7	small	red	metal	cylinder	-11,-26	8	medium	yellow	glass	cube	-23,21	9	small	blue	metal	cube	40,-23		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>2</td><td>position</td></tr> <tr><td>0</td><td>2</td><td>left,2</td></tr> </tbody> </table>	obj	attribute	value	0	2	position	0	2	left,2
obj	size	color	material	shape	position																																																																										
0	medium	yellow	glass	sphere	-34,10																																																																										
1	large	purple	metal	cylinder	-9,-8																																																																										
2	small	red	rubber	sphere	11,29																																																																										
3	small	yellow	glass	cube	-30,-20																																																																										
4	large	blue	rubber	cube	-13,35																																																																										
5	medium	yellow	glass	cylinder	-20,-37																																																																										
6	medium	red	rubber	sphere	17,-14																																																																										
7	small	red	metal	cylinder	-11,-26																																																																										
8	medium	yellow	glass	cube	-23,21																																																																										
9	small	blue	metal	cube	40,-23																																																																										
obj	attribute	value																																																																													
0	2	position																																																																													
0	2	left,2																																																																													
4		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>blue</td><td>glass</td><td>cube</td><td>-33,-21</td></tr> <tr><td>1</td><td>medium</td><td>cyan</td><td>metal</td><td>cube</td><td>12,11</td></tr> <tr><td>2</td><td>large</td><td>red</td><td>metal</td><td>cylinder</td><td>-16,23</td></tr> <tr><td>3</td><td>large</td><td>yellow</td><td>glass</td><td>cube</td><td>-28,12</td></tr> <tr><td>4</td><td>medium</td><td>red</td><td>rubber</td><td>cylinder</td><td>-11,7</td></tr> <tr><td>5</td><td>large</td><td>blue</td><td>metal</td><td>cylinder</td><td>16,-20</td></tr> <tr><td>6</td><td>medium</td><td>blue</td><td>glass</td><td>cylinder</td><td>-8,-16</td></tr> <tr><td>7</td><td>large</td><td>green</td><td>glass</td><td>sphere</td><td>9,0</td></tr> <tr><td>8</td><td>large</td><td>red</td><td>metal</td><td>cube</td><td>-4,-6</td></tr> <tr><td>9</td><td>small</td><td>gray</td><td>rubber</td><td>cube</td><td>18,3</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	blue	glass	cube	-33,-21	1	medium	cyan	metal	cube	12,11	2	large	red	metal	cylinder	-16,23	3	large	yellow	glass	cube	-28,12	4	medium	red	rubber	cylinder	-11,7	5	large	blue	metal	cylinder	16,-20	6	medium	blue	glass	cylinder	-8,-16	7	large	green	glass	sphere	9,0	8	large	red	metal	cube	-4,-6	9	small	gray	rubber	cube	18,3		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>1</td><td>shape</td></tr> <tr><td>0</td><td>1</td><td>sphere</td></tr> </tbody> </table>	obj	attribute	value	0	1	shape	0	1	sphere
obj	size	color	material	shape	position																																																																										
0	small	blue	glass	cube	-33,-21																																																																										
1	medium	cyan	metal	cube	12,11																																																																										
2	large	red	metal	cylinder	-16,23																																																																										
3	large	yellow	glass	cube	-28,12																																																																										
4	medium	red	rubber	cylinder	-11,7																																																																										
5	large	blue	metal	cylinder	16,-20																																																																										
6	medium	blue	glass	cylinder	-8,-16																																																																										
7	large	green	glass	sphere	9,0																																																																										
8	large	red	metal	cube	-4,-6																																																																										
9	small	gray	rubber	cube	18,3																																																																										
obj	attribute	value																																																																													
0	1	shape																																																																													
0	1	sphere																																																																													
5		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>red</td><td>glass</td><td>cylinder</td><td>3,-27</td></tr> <tr><td>1</td><td>large</td><td>brown</td><td>glass</td><td>cube</td><td>38,-39</td></tr> <tr><td>2</td><td>small</td><td>blue</td><td>rubber</td><td>sphere</td><td>36,-2</td></tr> <tr><td>3</td><td>large</td><td>red</td><td>metal</td><td>cube</td><td>10,-16</td></tr> <tr><td>4</td><td>large</td><td>green</td><td>metal</td><td>sphere</td><td>12,2</td></tr> <tr><td>5</td><td>small</td><td>brown</td><td>metal</td><td>sphere</td><td>-34,17</td></tr> <tr><td>6</td><td>small</td><td>green</td><td>rubber</td><td>cylinder</td><td>-12,15</td></tr> <tr><td>7</td><td>large</td><td>green</td><td>rubber</td><td>cylinder</td><td>-19,-19</td></tr> <tr><td>8</td><td>large</td><td>cyan</td><td>glass</td><td>sphere</td><td>-20,-1</td></tr> <tr><td>9</td><td>medium</td><td>green</td><td>metal</td><td>cube</td><td>-10,34</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	red	glass	cylinder	3,-27	1	large	brown	glass	cube	38,-39	2	small	blue	rubber	sphere	36,-2	3	large	red	metal	cube	10,-16	4	large	green	metal	sphere	12,2	5	small	brown	metal	sphere	-34,17	6	small	green	rubber	cylinder	-12,15	7	large	green	rubber	cylinder	-19,-19	8	large	cyan	glass	sphere	-20,-1	9	medium	green	metal	cube	-10,34		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>6</td><td>size</td></tr> <tr><td>0</td><td>6</td><td>large</td></tr> </tbody> </table>	obj	attribute	value	0	6	size	0	6	large
obj	size	color	material	shape	position																																																																										
0	medium	red	glass	cylinder	3,-27																																																																										
1	large	brown	glass	cube	38,-39																																																																										
2	small	blue	rubber	sphere	36,-2																																																																										
3	large	red	metal	cube	10,-16																																																																										
4	large	green	metal	sphere	12,2																																																																										
5	small	brown	metal	sphere	-34,17																																																																										
6	small	green	rubber	cylinder	-12,15																																																																										
7	large	green	rubber	cylinder	-19,-19																																																																										
8	large	cyan	glass	sphere	-20,-1																																																																										
9	medium	green	metal	cube	-10,34																																																																										
obj	attribute	value																																																																													
0	6	size																																																																													
0	6	large																																																																													

Figure 9. Examples from the Basic setting.

	Initial State (Top) Final State (Bottom)	Attributes of the Initial Objects	Visualization of the Initial Attributes	Reference Transformation																																																																																	
1		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>large</td><td>cyan</td><td>metal</td><td>sphere</td><td>3,20</td></tr> <tr><td>1</td><td>medium</td><td>blue</td><td>rubber</td><td>cube</td><td>19,-13</td></tr> <tr><td>2</td><td>medium</td><td>yellow</td><td>metal</td><td>cylinder</td><td>30,-31</td></tr> <tr><td>3</td><td>medium</td><td>cyan</td><td>rubber</td><td>cube</td><td>-18,6</td></tr> <tr><td>4</td><td>large</td><td>green</td><td>glass</td><td>sphere</td><td>-19,5</td></tr> <tr><td>5</td><td>medium</td><td>yellow</td><td>glass</td><td>sphere</td><td>30,0</td></tr> <tr><td>6</td><td>small</td><td>red</td><td>rubber</td><td>sphere</td><td>-9,-14</td></tr> <tr><td>7</td><td>small</td><td>gray</td><td>glass</td><td>cylinder</td><td>14,-36</td></tr> <tr><td>8</td><td>medium</td><td>brown</td><td>metal</td><td>cube</td><td>-15,-29</td></tr> <tr><td>9</td><td>medium</td><td>green</td><td>rubber</td><td>cylinder</td><td>14,8</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	large	cyan	metal	sphere	3,20	1	medium	blue	rubber	cube	19,-13	2	medium	yellow	metal	cylinder	30,-31	3	medium	cyan	rubber	cube	-18,6	4	large	green	glass	sphere	-19,5	5	medium	yellow	glass	sphere	30,0	6	small	red	rubber	sphere	-9,-14	7	small	gray	glass	cylinder	14,-36	8	medium	brown	metal	cube	-15,-29	9	medium	green	rubber	cylinder	14,8		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>position behind,1</td></tr> <tr><td>1</td><td>1</td><td>material metal</td></tr> <tr><td>2</td><td>9</td><td>color cyan</td></tr> <tr><td>3</td><td>4</td><td>position front,2</td></tr> </tbody> </table>	obj	attribute	value	0	0	position behind,1	1	1	material metal	2	9	color cyan	3	4	position front,2
obj	size	color	material	shape	position																																																																																
0	large	cyan	metal	sphere	3,20																																																																																
1	medium	blue	rubber	cube	19,-13																																																																																
2	medium	yellow	metal	cylinder	30,-31																																																																																
3	medium	cyan	rubber	cube	-18,6																																																																																
4	large	green	glass	sphere	-19,5																																																																																
5	medium	yellow	glass	sphere	30,0																																																																																
6	small	red	rubber	sphere	-9,-14																																																																																
7	small	gray	glass	cylinder	14,-36																																																																																
8	medium	brown	metal	cube	-15,-29																																																																																
9	medium	green	rubber	cylinder	14,8																																																																																
obj	attribute	value																																																																																			
0	0	position behind,1																																																																																			
1	1	material metal																																																																																			
2	9	color cyan																																																																																			
3	4	position front,2																																																																																			
2		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>gray</td><td>rubber</td><td>cube</td><td>18,31</td></tr> <tr><td>1</td><td>medium</td><td>green</td><td>rubber</td><td>cube</td><td>6,31</td></tr> <tr><td>2</td><td>medium</td><td>red</td><td>metal</td><td>cube</td><td>8,17</td></tr> <tr><td>3</td><td>medium</td><td>brown</td><td>rubber</td><td>sphere</td><td>-2,10</td></tr> <tr><td>4</td><td>large</td><td>cyan</td><td>metal</td><td>cube</td><td>12,0</td></tr> <tr><td>5</td><td>small</td><td>brown</td><td>rubber</td><td>sphere</td><td>-11,-4</td></tr> <tr><td>6</td><td>medium</td><td>green</td><td>rubber</td><td>cube</td><td>-20,-18</td></tr> <tr><td>7</td><td>small</td><td>red</td><td>metal</td><td>sphere</td><td>35,-40</td></tr> <tr><td>8</td><td>small</td><td>red</td><td>glass</td><td>cylinder</td><td>-34,21</td></tr> <tr><td>9</td><td>medium</td><td>gray</td><td>rubber</td><td>cube</td><td>20,16</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	gray	rubber	cube	18,31	1	medium	green	rubber	cube	6,31	2	medium	red	metal	cube	8,17	3	medium	brown	rubber	sphere	-2,10	4	large	cyan	metal	cube	12,0	5	small	brown	rubber	sphere	-11,-4	6	medium	green	rubber	cube	-20,-18	7	small	red	metal	sphere	35,-40	8	small	red	glass	cylinder	-34,21	9	medium	gray	rubber	cube	20,16		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>3</td><td>position front,1</td></tr> <tr><td>1</td><td>3</td><td>color gray</td></tr> <tr><td>2</td><td>6</td><td>position left,2</td></tr> <tr><td>3</td><td>3</td><td>position right,1</td></tr> </tbody> </table>	obj	attribute	value	0	3	position front,1	1	3	color gray	2	6	position left,2	3	3	position right,1
obj	size	color	material	shape	position																																																																																
0	small	gray	rubber	cube	18,31																																																																																
1	medium	green	rubber	cube	6,31																																																																																
2	medium	red	metal	cube	8,17																																																																																
3	medium	brown	rubber	sphere	-2,10																																																																																
4	large	cyan	metal	cube	12,0																																																																																
5	small	brown	rubber	sphere	-11,-4																																																																																
6	medium	green	rubber	cube	-20,-18																																																																																
7	small	red	metal	sphere	35,-40																																																																																
8	small	red	glass	cylinder	-34,21																																																																																
9	medium	gray	rubber	cube	20,16																																																																																
obj	attribute	value																																																																																			
0	3	position front,1																																																																																			
1	3	color gray																																																																																			
2	6	position left,2																																																																																			
3	3	position right,1																																																																																			
3		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>yellow</td><td>glass</td><td>sphere</td><td>-2,17</td></tr> <tr><td>1</td><td>medium</td><td>gray</td><td>rubber</td><td>cube</td><td>-14,16</td></tr> <tr><td>2</td><td>small</td><td>green</td><td>metal</td><td>cube</td><td>-39,-31</td></tr> <tr><td>3</td><td>medium</td><td>brown</td><td>glass</td><td>cube</td><td>-6,-31</td></tr> <tr><td>4</td><td>small</td><td>cyan</td><td>glass</td><td>cylinder</td><td>-4,-1</td></tr> <tr><td>5</td><td>medium</td><td>green</td><td>metal</td><td>cylinder</td><td>33,-40</td></tr> <tr><td>6</td><td>medium</td><td>green</td><td>rubber</td><td>sphere</td><td>-8,-22</td></tr> <tr><td>7</td><td>large</td><td>cyan</td><td>rubber</td><td>sphere</td><td>15,15</td></tr> <tr><td>8</td><td>small</td><td>blue</td><td>glass</td><td>cube</td><td>-2,-10</td></tr> <tr><td>9</td><td>small</td><td>brown</td><td>rubber</td><td>cube</td><td>18,-20</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	yellow	glass	sphere	-2,17	1	medium	gray	rubber	cube	-14,16	2	small	green	metal	cube	-39,-31	3	medium	brown	glass	cube	-6,-31	4	small	cyan	glass	cylinder	-4,-1	5	medium	green	metal	cylinder	33,-40	6	medium	green	rubber	sphere	-8,-22	7	large	cyan	rubber	sphere	15,15	8	small	blue	glass	cube	-2,-10	9	small	brown	rubber	cube	18,-20		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>0</td><td>material rubber</td></tr> <tr><td>1</td><td>4</td><td>color blue</td></tr> <tr><td>2</td><td>1</td><td>material glass</td></tr> </tbody> </table>	obj	attribute	value	0	0	material rubber	1	4	color blue	2	1	material glass			
obj	size	color	material	shape	position																																																																																
0	small	yellow	glass	sphere	-2,17																																																																																
1	medium	gray	rubber	cube	-14,16																																																																																
2	small	green	metal	cube	-39,-31																																																																																
3	medium	brown	glass	cube	-6,-31																																																																																
4	small	cyan	glass	cylinder	-4,-1																																																																																
5	medium	green	metal	cylinder	33,-40																																																																																
6	medium	green	rubber	sphere	-8,-22																																																																																
7	large	cyan	rubber	sphere	15,15																																																																																
8	small	blue	glass	cube	-2,-10																																																																																
9	small	brown	rubber	cube	18,-20																																																																																
obj	attribute	value																																																																																			
0	0	material rubber																																																																																			
1	4	color blue																																																																																			
2	1	material glass																																																																																			
4		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>purple</td><td>glass</td><td>cylinder</td><td>3,-20</td></tr> <tr><td>1</td><td>medium</td><td>gray</td><td>metal</td><td>sphere</td><td>-9,-18</td></tr> <tr><td>2</td><td>large</td><td>yellow</td><td>rubber</td><td>sphere</td><td>-37,-38</td></tr> <tr><td>3</td><td>large</td><td>cyan</td><td>glass</td><td>sphere</td><td>-4,-10</td></tr> <tr><td>4</td><td>medium</td><td>green</td><td>rubber</td><td>cube</td><td>38,39</td></tr> <tr><td>5</td><td>medium</td><td>red</td><td>rubber</td><td>sphere</td><td>13,-17</td></tr> <tr><td>6</td><td>large</td><td>gray</td><td>glass</td><td>cube</td><td>14,15</td></tr> <tr><td>7</td><td>large</td><td>blue</td><td>glass</td><td>cylinder</td><td>-12,1</td></tr> <tr><td>8</td><td>small</td><td>yellow</td><td>rubber</td><td>sphere</td><td>-23,22</td></tr> <tr><td>9</td><td>small</td><td>gray</td><td>rubber</td><td>cylinder</td><td>1,32</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	purple	glass	cylinder	3,-20	1	medium	gray	metal	sphere	-9,-18	2	large	yellow	rubber	sphere	-37,-38	3	large	cyan	glass	sphere	-4,-10	4	medium	green	rubber	cube	38,39	5	medium	red	rubber	sphere	13,-17	6	large	gray	glass	cube	14,15	7	large	blue	glass	cylinder	-12,1	8	small	yellow	rubber	sphere	-23,22	9	small	gray	rubber	cylinder	1,32		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>6</td><td>size small</td></tr> <tr><td>1</td><td>3</td><td>color brown</td></tr> <tr><td>2</td><td>7</td><td>color green</td></tr> </tbody> </table>	obj	attribute	value	0	6	size small	1	3	color brown	2	7	color green			
obj	size	color	material	shape	position																																																																																
0	small	purple	glass	cylinder	3,-20																																																																																
1	medium	gray	metal	sphere	-9,-18																																																																																
2	large	yellow	rubber	sphere	-37,-38																																																																																
3	large	cyan	glass	sphere	-4,-10																																																																																
4	medium	green	rubber	cube	38,39																																																																																
5	medium	red	rubber	sphere	13,-17																																																																																
6	large	gray	glass	cube	14,15																																																																																
7	large	blue	glass	cylinder	-12,1																																																																																
8	small	yellow	rubber	sphere	-23,22																																																																																
9	small	gray	rubber	cylinder	1,32																																																																																
obj	attribute	value																																																																																			
0	6	size small																																																																																			
1	3	color brown																																																																																			
2	7	color green																																																																																			
5		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>yellow</td><td>metal</td><td>sphere</td><td>15,18</td></tr> <tr><td>1</td><td>small</td><td>brown</td><td>glass</td><td>cube</td><td>31,10</td></tr> <tr><td>2</td><td>large</td><td>blue</td><td>glass</td><td>cylinder</td><td>8,-3</td></tr> <tr><td>3</td><td>medium</td><td>cyan</td><td>glass</td><td>cube</td><td>32,-18</td></tr> <tr><td>4</td><td>medium</td><td>brown</td><td>rubber</td><td>cylinder</td><td>6,14</td></tr> <tr><td>5</td><td>small</td><td>purple</td><td>metal</td><td>sphere</td><td>-14,10</td></tr> <tr><td>6</td><td>medium</td><td>red</td><td>glass</td><td>cube</td><td>-38,-39</td></tr> <tr><td>7</td><td>small</td><td>brown</td><td>metal</td><td>cube</td><td>2,7</td></tr> <tr><td>8</td><td>medium</td><td>yellow</td><td>metal</td><td>cube</td><td>-33,16</td></tr> <tr><td>9</td><td>medium</td><td>yellow</td><td>rubber</td><td>cube</td><td>-34,27</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	yellow	metal	sphere	15,18	1	small	brown	glass	cube	31,10	2	large	blue	glass	cylinder	8,-3	3	medium	cyan	glass	cube	32,-18	4	medium	brown	rubber	cylinder	6,14	5	small	purple	metal	sphere	-14,10	6	medium	red	glass	cube	-38,-39	7	small	brown	metal	cube	2,7	8	medium	yellow	metal	cube	-33,16	9	medium	yellow	rubber	cube	-34,27		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>3</td><td>position front,2</td></tr> <tr><td>1</td><td>0</td><td>position right,1</td></tr> <tr><td>2</td><td>0</td><td>position front-left,2</td></tr> <tr><td>3</td><td>2</td><td>shape sphere</td></tr> </tbody> </table>	obj	attribute	value	0	3	position front,2	1	0	position right,1	2	0	position front-left,2	3	2	shape sphere
obj	size	color	material	shape	position																																																																																
0	medium	yellow	metal	sphere	15,18																																																																																
1	small	brown	glass	cube	31,10																																																																																
2	large	blue	glass	cylinder	8,-3																																																																																
3	medium	cyan	glass	cube	32,-18																																																																																
4	medium	brown	rubber	cylinder	6,14																																																																																
5	small	purple	metal	sphere	-14,10																																																																																
6	medium	red	glass	cube	-38,-39																																																																																
7	small	brown	metal	cube	2,7																																																																																
8	medium	yellow	metal	cube	-33,16																																																																																
9	medium	yellow	rubber	cube	-34,27																																																																																
obj	attribute	value																																																																																			
0	3	position front,2																																																																																			
1	0	position right,1																																																																																			
2	0	position front-left,2																																																																																			
3	2	shape sphere																																																																																			

Figure 10. Examples from the Event setting.

	Initial State (Top) Final State (Bottom)	Attributes of the Initial Objects	Visualization of the Initial Attributes	Reference Transformation																																																																																	
1		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>red</td><td>rubber</td><td>cube</td><td>16,-20</td></tr> <tr><td>1</td><td>medium</td><td>red</td><td>rubber</td><td>cylinder</td><td>-6,17</td></tr> <tr><td>2</td><td>small</td><td>green</td><td>glass</td><td>sphere</td><td>2,19</td></tr> <tr><td>3</td><td>small</td><td>gray</td><td>rubber</td><td>cube</td><td>24,23</td></tr> <tr><td>4</td><td>large</td><td>blue</td><td>rubber</td><td>cube</td><td>-18,16</td></tr> <tr><td>5</td><td>large</td><td>red</td><td>metal</td><td>cylinder</td><td>20,-4</td></tr> <tr><td>6</td><td>small</td><td>green</td><td>metal</td><td>sphere</td><td>-8,-29</td></tr> <tr><td>7</td><td>small</td><td>red</td><td>metal</td><td>cylinder</td><td>-20,-8</td></tr> <tr><td>8</td><td>medium</td><td>blue</td><td>metal</td><td>cube</td><td>2,-30</td></tr> <tr><td>9</td><td>large</td><td>blue</td><td>glass</td><td>cube</td><td>23,-31</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	red	rubber	cube	16,-20	1	medium	red	rubber	cylinder	-6,17	2	small	green	glass	sphere	2,19	3	small	gray	rubber	cube	24,23	4	large	blue	rubber	cube	-18,16	5	large	red	metal	cylinder	20,-4	6	small	green	metal	sphere	-8,-29	7	small	red	metal	cylinder	-20,-8	8	medium	blue	metal	cube	2,-30	9	large	blue	glass	cube	23,-31		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>5</td><td>position front,1</td></tr> <tr><td>1</td><td>5</td><td>color cyan</td></tr> <tr><td>2</td><td>0</td><td>size medium</td></tr> <tr><td>3</td><td>1</td><td>position right,2</td></tr> </tbody> </table>	obj	attribute	value	0	5	position front,1	1	5	color cyan	2	0	size medium	3	1	position right,2
obj	size	color	material	shape	position																																																																																
0	small	red	rubber	cube	16,-20																																																																																
1	medium	red	rubber	cylinder	-6,17																																																																																
2	small	green	glass	sphere	2,19																																																																																
3	small	gray	rubber	cube	24,23																																																																																
4	large	blue	rubber	cube	-18,16																																																																																
5	large	red	metal	cylinder	20,-4																																																																																
6	small	green	metal	sphere	-8,-29																																																																																
7	small	red	metal	cylinder	-20,-8																																																																																
8	medium	blue	metal	cube	2,-30																																																																																
9	large	blue	glass	cube	23,-31																																																																																
obj	attribute	value																																																																																			
0	5	position front,1																																																																																			
1	5	color cyan																																																																																			
2	0	size medium																																																																																			
3	1	position right,2																																																																																			
2		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>small</td><td>yellow</td><td>metal</td><td>cylinder</td><td>21,-38</td></tr> <tr><td>1</td><td>large</td><td>yellow</td><td>rubber</td><td>cube</td><td>-4,1</td></tr> <tr><td>2</td><td>large</td><td>green</td><td>metal</td><td>sphere</td><td>-9,20</td></tr> <tr><td>3</td><td>medium</td><td>purple</td><td>rubber</td><td>sphere</td><td>38,-21</td></tr> <tr><td>4</td><td>medium</td><td>red</td><td>rubber</td><td>cylinder</td><td>13,-11</td></tr> <tr><td>5</td><td>medium</td><td>blue</td><td>rubber</td><td>sphere</td><td>12,9</td></tr> <tr><td>6</td><td>large</td><td>green</td><td>rubber</td><td>cube</td><td>-3,-14</td></tr> <tr><td>7</td><td>large</td><td>blue</td><td>rubber</td><td>sphere</td><td>-3,-31</td></tr> <tr><td>8</td><td>medium</td><td>green</td><td>metal</td><td>sphere</td><td>-21,-16</td></tr> <tr><td>9</td><td>small</td><td>cyan</td><td>rubber</td><td>cylinder</td><td>-23,7</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	small	yellow	metal	cylinder	21,-38	1	large	yellow	rubber	cube	-4,1	2	large	green	metal	sphere	-9,20	3	medium	purple	rubber	sphere	38,-21	4	medium	red	rubber	cylinder	13,-11	5	medium	blue	rubber	sphere	12,9	6	large	green	rubber	cube	-3,-14	7	large	blue	rubber	sphere	-3,-31	8	medium	green	metal	sphere	-21,-16	9	small	cyan	rubber	cylinder	-23,7		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>5</td><td>position behind-left,2</td></tr> <tr><td>1</td><td>1</td><td>color green</td></tr> <tr><td>2</td><td>2</td><td>size medium</td></tr> <tr><td>3</td><td>2</td><td>position front-left,1</td></tr> </tbody> </table>	obj	attribute	value	0	5	position behind-left,2	1	1	color green	2	2	size medium	3	2	position front-left,1
obj	size	color	material	shape	position																																																																																
0	small	yellow	metal	cylinder	21,-38																																																																																
1	large	yellow	rubber	cube	-4,1																																																																																
2	large	green	metal	sphere	-9,20																																																																																
3	medium	purple	rubber	sphere	38,-21																																																																																
4	medium	red	rubber	cylinder	13,-11																																																																																
5	medium	blue	rubber	sphere	12,9																																																																																
6	large	green	rubber	cube	-3,-14																																																																																
7	large	blue	rubber	sphere	-3,-31																																																																																
8	medium	green	metal	sphere	-21,-16																																																																																
9	small	cyan	rubber	cylinder	-23,7																																																																																
obj	attribute	value																																																																																			
0	5	position behind-left,2																																																																																			
1	1	color green																																																																																			
2	2	size medium																																																																																			
3	2	position front-left,1																																																																																			
3		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>brown</td><td>glass</td><td>sphere</td><td>-29,38</td></tr> <tr><td>1</td><td>medium</td><td>cyan</td><td>glass</td><td>cube</td><td>-17,14</td></tr> <tr><td>2</td><td>large</td><td>gray</td><td>metal</td><td>cylinder</td><td>-27,-11</td></tr> <tr><td>3</td><td>small</td><td>purple</td><td>glass</td><td>cylinder</td><td>13,19</td></tr> <tr><td>4</td><td>medium</td><td>purple</td><td>rubber</td><td>cube</td><td>20,-13</td></tr> <tr><td>5</td><td>large</td><td>red</td><td>rubber</td><td>cube</td><td>0,16</td></tr> <tr><td>6</td><td>medium</td><td>gray</td><td>metal</td><td>cylinder</td><td>-15,-5</td></tr> <tr><td>7</td><td>small</td><td>blue</td><td>glass</td><td>cylinder</td><td>27,37</td></tr> <tr><td>8</td><td>small</td><td>green</td><td>metal</td><td>sphere</td><td>-6,3</td></tr> <tr><td>9</td><td>small</td><td>blue</td><td>rubber</td><td>cylinder</td><td>38,12</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	brown	glass	sphere	-29,38	1	medium	cyan	glass	cube	-17,14	2	large	gray	metal	cylinder	-27,-11	3	small	purple	glass	cylinder	13,19	4	medium	purple	rubber	cube	20,-13	5	large	red	rubber	cube	0,16	6	medium	gray	metal	cylinder	-15,-5	7	small	blue	glass	cylinder	27,37	8	small	green	metal	sphere	-6,3	9	small	blue	rubber	cylinder	38,12		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>8</td><td>material rubber</td></tr> <tr><td>1</td><td>4</td><td>size small</td></tr> <tr><td>2</td><td>5</td><td>shape cylinder</td></tr> </tbody> </table>	obj	attribute	value	0	8	material rubber	1	4	size small	2	5	shape cylinder			
obj	size	color	material	shape	position																																																																																
0	medium	brown	glass	sphere	-29,38																																																																																
1	medium	cyan	glass	cube	-17,14																																																																																
2	large	gray	metal	cylinder	-27,-11																																																																																
3	small	purple	glass	cylinder	13,19																																																																																
4	medium	purple	rubber	cube	20,-13																																																																																
5	large	red	rubber	cube	0,16																																																																																
6	medium	gray	metal	cylinder	-15,-5																																																																																
7	small	blue	glass	cylinder	27,37																																																																																
8	small	green	metal	sphere	-6,3																																																																																
9	small	blue	rubber	cylinder	38,12																																																																																
obj	attribute	value																																																																																			
0	8	material rubber																																																																																			
1	4	size small																																																																																			
2	5	shape cylinder																																																																																			
4		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>cyan</td><td>glass</td><td>cube</td><td>32,21</td></tr> <tr><td>1</td><td>large</td><td>blue</td><td>rubber</td><td>cube</td><td>11,-1</td></tr> <tr><td>2</td><td>medium</td><td>green</td><td>metal</td><td>sphere</td><td>17,-27</td></tr> <tr><td>3</td><td>large</td><td>blue</td><td>metal</td><td>cylinder</td><td>-22,-29</td></tr> <tr><td>4</td><td>medium</td><td>green</td><td>glass</td><td>cube</td><td>-27,13</td></tr> <tr><td>5</td><td>large</td><td>red</td><td>metal</td><td>cube</td><td>1,16</td></tr> <tr><td>6</td><td>medium</td><td>brown</td><td>glass</td><td>cylinder</td><td>7,-38</td></tr> <tr><td>7</td><td>medium</td><td>purple</td><td>metal</td><td>sphere</td><td>-20,-16</td></tr> <tr><td>8</td><td>medium</td><td>cyan</td><td>metal</td><td>sphere</td><td>3,-15</td></tr> <tr><td>9</td><td>large</td><td>cyan</td><td>metal</td><td>cube</td><td>-8,1</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	cyan	glass	cube	32,21	1	large	blue	rubber	cube	11,-1	2	medium	green	metal	sphere	17,-27	3	large	blue	metal	cylinder	-22,-29	4	medium	green	glass	cube	-27,13	5	large	red	metal	cube	1,16	6	medium	brown	glass	cylinder	7,-38	7	medium	purple	metal	sphere	-20,-16	8	medium	cyan	metal	sphere	3,-15	9	large	cyan	metal	cube	-8,1		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>4</td><td>position behind,1</td></tr> <tr><td>1</td><td>5</td><td>size medium</td></tr> <tr><td>2</td><td>7</td><td>shape cube</td></tr> <tr><td>3</td><td>7</td><td>material glass</td></tr> </tbody> </table>	obj	attribute	value	0	4	position behind,1	1	5	size medium	2	7	shape cube	3	7	material glass
obj	size	color	material	shape	position																																																																																
0	medium	cyan	glass	cube	32,21																																																																																
1	large	blue	rubber	cube	11,-1																																																																																
2	medium	green	metal	sphere	17,-27																																																																																
3	large	blue	metal	cylinder	-22,-29																																																																																
4	medium	green	glass	cube	-27,13																																																																																
5	large	red	metal	cube	1,16																																																																																
6	medium	brown	glass	cylinder	7,-38																																																																																
7	medium	purple	metal	sphere	-20,-16																																																																																
8	medium	cyan	metal	sphere	3,-15																																																																																
9	large	cyan	metal	cube	-8,1																																																																																
obj	attribute	value																																																																																			
0	4	position behind,1																																																																																			
1	5	size medium																																																																																			
2	7	shape cube																																																																																			
3	7	material glass																																																																																			
5		<table border="1"> <thead> <tr> <th>obj</th><th>size</th><th>color</th><th>material</th><th>shape</th><th>position</th></tr> </thead> <tbody> <tr><td>0</td><td>medium</td><td>red</td><td>rubber</td><td>sphere</td><td>-14,-1</td></tr> <tr><td>1</td><td>small</td><td>green</td><td>metal</td><td>cube</td><td>12,-1</td></tr> <tr><td>2</td><td>small</td><td>gray</td><td>glass</td><td>sphere</td><td>-34,9</td></tr> <tr><td>3</td><td>large</td><td>cyan</td><td>glass</td><td>cube</td><td>-33,-32</td></tr> <tr><td>4</td><td>medium</td><td>blue</td><td>glass</td><td>cube</td><td>15,-10</td></tr> <tr><td>5</td><td>small</td><td>yellow</td><td>rubber</td><td>cylinder</td><td>-29,20</td></tr> <tr><td>6</td><td>small</td><td>blue</td><td>metal</td><td>sphere</td><td>36,29</td></tr> <tr><td>7</td><td>medium</td><td>purple</td><td>glass</td><td>cube</td><td>16,21</td></tr> <tr><td>8</td><td>medium</td><td>cyan</td><td>metal</td><td>sphere</td><td>28,-36</td></tr> <tr><td>9</td><td>large</td><td>purple</td><td>rubber</td><td>cylinder</td><td>-17,10</td></tr> </tbody> </table>	obj	size	color	material	shape	position	0	medium	red	rubber	sphere	-14,-1	1	small	green	metal	cube	12,-1	2	small	gray	glass	sphere	-34,9	3	large	cyan	glass	cube	-33,-32	4	medium	blue	glass	cube	15,-10	5	small	yellow	rubber	cylinder	-29,20	6	small	blue	metal	sphere	36,29	7	medium	purple	glass	cube	16,21	8	medium	cyan	metal	sphere	28,-36	9	large	purple	rubber	cylinder	-17,10		<table border="1"> <thead> <tr> <th>obj</th><th>attribute</th><th>value</th></tr> </thead> <tbody> <tr><td>0</td><td>4</td><td>position right,2</td></tr> <tr><td>1</td><td>9</td><td>material metal</td></tr> <tr><td>2</td><td>9</td><td>color gray</td></tr> </tbody> </table>	obj	attribute	value	0	4	position right,2	1	9	material metal	2	9	color gray			
obj	size	color	material	shape	position																																																																																
0	medium	red	rubber	sphere	-14,-1																																																																																
1	small	green	metal	cube	12,-1																																																																																
2	small	gray	glass	sphere	-34,9																																																																																
3	large	cyan	glass	cube	-33,-32																																																																																
4	medium	blue	glass	cube	15,-10																																																																																
5	small	yellow	rubber	cylinder	-29,20																																																																																
6	small	blue	metal	sphere	36,29																																																																																
7	medium	purple	glass	cube	16,21																																																																																
8	medium	cyan	metal	sphere	28,-36																																																																																
9	large	purple	rubber	cylinder	-17,10																																																																																
obj	attribute	value																																																																																			
0	4	position right,2																																																																																			
1	9	material metal																																																																																			
2	9	color gray																																																																																			

Figure 11. Examples from the View setting.

## References

- [1] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2127–2136, 2017. 3
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2
- [3] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, pages 5082–5093, 2019. 2
- [4] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. In *ICLR*, 2020. 2
- [5] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. *arXiv preprint arXiv:1907.01172*, 2019. 3
- [6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 2, 6
- [7] Blender Online Community. Blender—a 3d modelling and rendering package, 2016. 3
- [8] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2586, 2013. 3
- [9] Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*, 2020. 2
- [10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 6, 10
- [13] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016. 2
- [14] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 2, 5
- [15] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2
- [16] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 3
- [17] Jung-Hwan Jin, Hyun Joon Shin, and Jung-Ju Choi. Spoid: a system to produce spot-the-difference puzzle images with difficulty. *The Visual Computer*, 29(6-8):481–489, 2013. 5
- [18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1, 2, 3
- [19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017. 2, 5
- [20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 10
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 10
- [22] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020. 3
- [23] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 2, 6, 10
- [24] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019. 4
- [25] Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2924–2932, 2017. 3
- [26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. 2
- [27] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 3
- [28] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 6

- [29] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 9
- [31] Jean Piaget. The role of action in the development of thinking. In *Knowledge and development*, pages 17–42. Springer, 1977. 2
- [32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1
- [33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1
- [34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 10
- [35] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 1, 2
- [36] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019. 1, 2
- [37] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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 4149–4158, 2019. 2
- [38] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. FVQA: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018. 2
- [39] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions~transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016. 3
- [40] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 7
- [41] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989. 6
- [42] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*, 2018. 2
- [43] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 2
- [44] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Ji-jun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. 2
- [45] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2
- [46] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [47] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 2
- [48] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 521–529, 2019. 3