A. Details of compositional reasoning frameworks

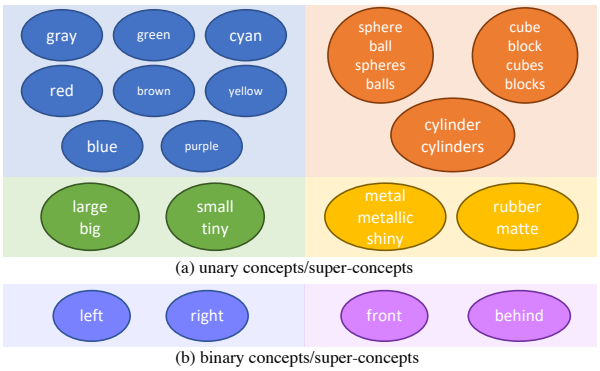
Baseline visual reasoning framework The original compositional reasoning framework [19] is similar to the phase 1 of our framework in Figure 2 of the main paper, except that it works on pixel-level instead of object-level features. To generate vs, it feeds the image to a ResNet101 [14] pretrained on ImageNet [10] and flatten the last feature maps across the width and height as vs. For the question inputs, we first convert each question word to its word embedding vector (ws), then input ws to a bidirectional LSTM [15, 11] to extract the question embedding vector q. The compositional reasoning module takes vs, ws and q as inputs and performs multi-step reasoning to attain m, the final step memory output. Finally, the classifier outputs the probability for each answer choice with a linear classifier over the concatenation of m and q.

The MAC reasoning module At each step, the i-th MAC cell receives the control signal c_{i-1} and the memory output from the previous step, m_{i-1}, and outputs the new memory vector m_i. The control unit computes the single c_i to control reading of vs in the R/W unit. Specifically, it computes the interactions among c_{i-1}, q_i, and each vector in ws to produce the attention weights, and weighted averages ws to produce c_i. The control unit of each MAC cell has a unique question embedding projection layer, while all other layers are shared. The R/W unit aims to read the useful vs and store the read information into m_i. It first computes the interactions among m_{i-1}, c_{i-1} and each vector in vs to attain the attention weights, weighted averages vs to produce a read vector r_i, and finally computes the interaction of r_i and m_{i-1} to produce m_i. The weights of the R/W units are shared across all MAC cells. The initial control signal and memory c_0 and m_0 are learnable parameters.

B. Implementation details

CLEVR We set the hidden dimension D to 512 in all modules. We follow [19] to design the question embedding module, the compositional module and the classifier. For the object-level feature extracter, we make the backbone ResNet34 learnable and zero-pad the output vs to 12 vectors in total for any image. Notice that the maximum number of objects in an image is 11, so that the reasoning module is able to read nothing into the memory for some steps. For the concept projection module, to cover the full view of vs, the conv1D consists of five 1D convolution layers with kernel sizes (7,5,5,5,5), each followed by a Batch Norm layer [22] and an ELU activation layer [9].

We use Adam optimizer [28] with momentum 0.9 and 0.999. Phase 1 and phase 2 share a same training schedule: the learning rate is initiated with 10^{-4} for the first 20 epochs and is halved every 5 epochs afterwards until stopped at the 40th epoch. We train the concept regression module separately with learning rate of 10^{-4} for 6 epochs. All the training process is conducted with a batch size of 256.



(a) unary concepts/super-concepts

(b) binary concepts/super-concepts

Figure 10: Concepts and super concept sets. Each circle represents a concept described by the words in that circle. A super concept set comprises the concepts represented by circles of the same color.

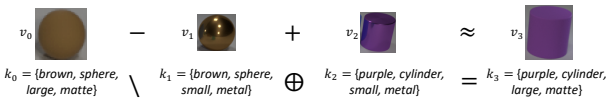


Figure 11: An multi-modal analogy example enabled by our results.

GQA The implementation details in the GQA setting basically follows the details on CLEVR. To better handle the complexity in GQA, we concatenate the object features with their corresponding bounding box coordinates to enhance the objects' location representations similar to [18]. We use GloVe [36] to initialize question word embeddings and maintain an exponential moving average with a decay rate of 0.999 to update the model parameters.

C. Visualization of the induced concept hierarchy

After visual mapping, binary coding and concept/super-concept induction, the unary concepts and super concepts are induced as shown in Figure 10; the binary concepts are 'left', 'right', 'front' and 'behind', and {'left', 'right'} and {'front', 'behind'} form two super concept sets. The generated concept hierarchy perfectly recovers the definition in CLEVR data generator and matches human prior knowledge, showing the success of our approach.

D. Multi-modal concept analogy

Our concept induction results bridge the visual and symbolic spaces. The results enable to extend word anal-

ogy [34] (e.g., “Madrid” - “Spain” + “France” $\rightarrow$ “Paris”) into the multi-modality setting. Figure 11 gives an example, starting with the initial object v_0 and its predicted concepts K_0, subtracting concepts K_1 and adding new concepts K_2 result in a new concept set K_3 (Figure 11 (bottom)). Then if we retrieve visual object v_i with each concept set K_i along the path (Figure 11 (top)), we have $v_0 - v_1 + v_2 \approx v_3$ in the original visual feature space.

E. Derivation from the concept interpretation

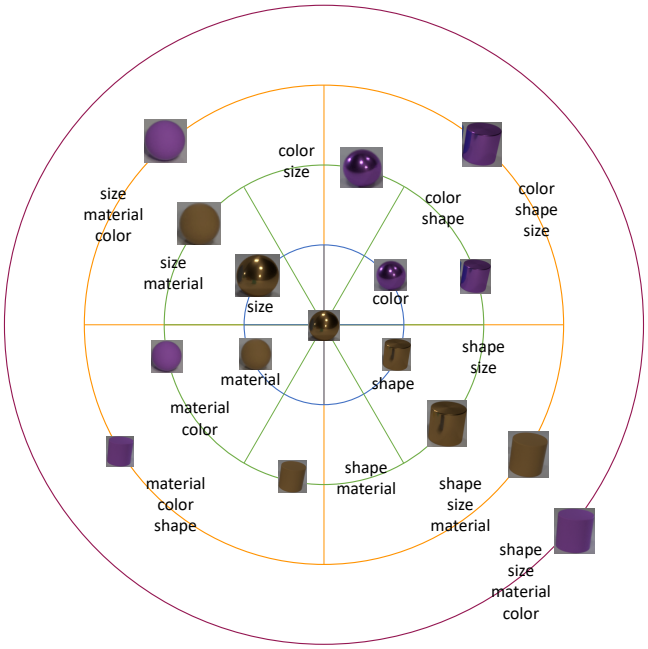


Figure 12: Illustration of the semantic distance.

With the induced concepts and super concept sets, each object can be represented with a zero-one vector, k, where the entry is 1 if that object possesses the corresponding concept or 0 otherwise. Notice that the super concept sets split the whole concept set; we thereby name the entries of k corresponding to one super concept set as a super concept. The super concept is thus a zero-one vector with exactly one entry to be 1. We name this pattern as the super concept constraint. Therefore, we can define the semantic distance between two visual objects by the number of different super concepts or by Eqn. (5).

$$\zeta_{k_1,k_2} = \frac{|k_1 \oplus k_2|_1^1}{2}, \tag{5}$$

where k_1 and k_2 are the concept vectors representing two objects and $\oplus$ is the operation XOR. Studying the concepts and super concept sets induced, we acknowledge that the super concept sets correspond to color, shape, size and material in semantics. Thereby, we give an example of the semantic distances of multiple objects to one object as shown in Figure 12. The circle radii indicate the semantic distances to the object at the centers of these circles. The inner three circles are segmented so that each segment represents what super concepts are different. The outer circle represents all the 4 super concepts are different between the object on that circle and the object at the center.

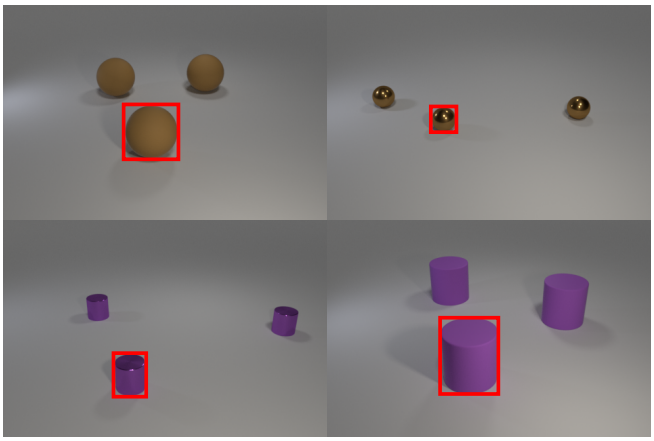


Figure 13: The original images for extracting visual features. The object-level features corresponding to the objects bounded by red rectangles are used for the illustration of semantic operations in the visual feature space.

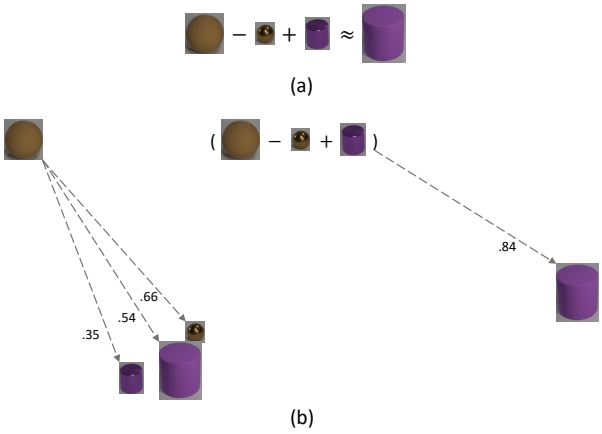


Figure 14: Illustration of the semantic analogy in the visual feature space. (a) The operations on the visual features. (b) The cosine similarities between pairs of visual feature vectors.

We can further interpret the semantic analogy in the visual feature space with the induced concept vectors. Shown in Figure 13, we first generate four images of different objects; then, we use our trained OCCAM structure to extract the object-level features corresponding to the objects bounded by red rectangles. Shown in Figure 14(a), we can move the visual feature vector of the leftmost object closer

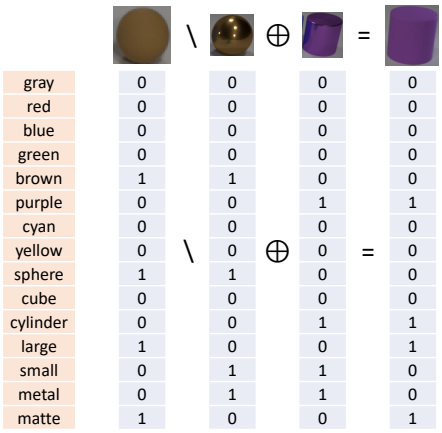


Figure 15: Operations on the concept vectors.

to that of the rightmost object by subtracting and adding visual feature vectors of two other objects. The proximity between pairs of visual feature vectors is measured with cosine similarity as shown in Figure 14(b). In the concept vector space, we can define a 'minus' operation, $k_1 \backslash k_2$, as eliminate the shared super concepts between k_1 and k_2 from k_1. We can also define a 'plus' operation, $k_1' \oplus k_2$, between a concept vector template k_1' and a concept vector k_2 as add the super concepts of o_2 that o_1' misses to o_1'. Therefore, The operations in the visual feature space can be explained with the operations we defined in the concept vector space shown in Figure 15.

F. Visualization of reasoning steps

We give an example of the compositional reasoning steps on the induced concept space of OCCAM as shown in Figure 16. While the attention is directly imposed on the projected concept vectors in the read unit of the compositional reasoning module, the attention can be equally mapped to the concept vectors and the visual objects as the projected concept vector to the concept vector or the projected concept vector to the visual object is a one-to-one mapping relationship. We also give an example of the compositional reasoning steps on the GQA dataset shown in Figure 17. As the dimension of the induced concept vectors is too high, here we only present the attention on objects in the image.

G. Human study

We assess the concept and super concept induction by studying how the word correlation conforms with our human knowledge. We present an extended subset of GQA concept correlations shown in Figure 18. It consists of the 98 most common single words for describing objects. Each entry in the matrix represents the conditional probability that the column attribute exists given the row attribute exists. A pair of mutual high correlation values between two words indicates that these words belong to the same concept, while the opposite means that the concepts represented by those words belong to a super concept. Therefore, we can evaluate the concept induction by assessing the conditional probabilities of synonyms or uncorrelated words for each word, because from us human understanding, a synonym is used to describe the same concept while an uncorrelated word describes a concept belonging to the same super concept.

For each word in the extended subset words, we first let annotators choose 2 synonyms and 2 uncorrelated words from the rest 97 words. Then, rank the four chosen words in a descending order of similarity between them and the original word. Based on these annotations, we conduct two experiments: 1) measure the accuracy of classifying the chosen words to synonyms and uncorrelated words; 2) measure the Kendall tau distance [26] between the word similarity ranking based on the conditional probability and that ranking based on human knowledge.

For the first experiment, we use a binary classifier with threshold 0.5 to classify the chosen words by humans. If a word's conditional probability given the original word is greater than the threshold, this word is classified as a synonym; if smaller, this word is classified as an uncorrelated word. The accuracy can be calculated with Eqn. (6).

$$A = \frac{1}{|S|}\sum_{i\in S}\frac{1}{|W_i|}\Big(\sum_{j\in W_i^{pos}}\mathbb{1}(R_{i,j} > t) + \sum_{j\in W_i^{neg}}\mathbb{1}(R_{i,j} < t)\Big), \tag{6}$$

where A represents accuracy, S is the subset of words, W_i represents the set of synonyms and uncorrelated words chosen for word i, $R_{i,j}$ represents the conditional probability of word j given word i exists and t is the threshold. For comparison, we also calculate the cosine similarity of word GloVe [36] embeddings to substitute the conditional probability and serve as R in Eqn. (6). For this setting, we tune the threshold t to be 0.21 to reach the best accuracy. The result in Table 4 shows that our induction highly conforms with our human sense in grouping words into concepts but does not agree much with humans in grouping super concepts. By further studying specific cases, we realize that a word and its uncorrelated words defined by humans can simultaneously describe one object. For example, 'white' and 'black' can be used together to describe a zebra; 'leafy' and 'leafless' both describes a status of a plant. Such words have high correlations, which defects with our human understanding.

The second experiment measures how the induced word proximity conforms with our human knowledge. For a word w_i, our annotators rank the chosen synonyms and uncorrelated words $a_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$ with a descending order of word similarity to w_i and assign a sequence

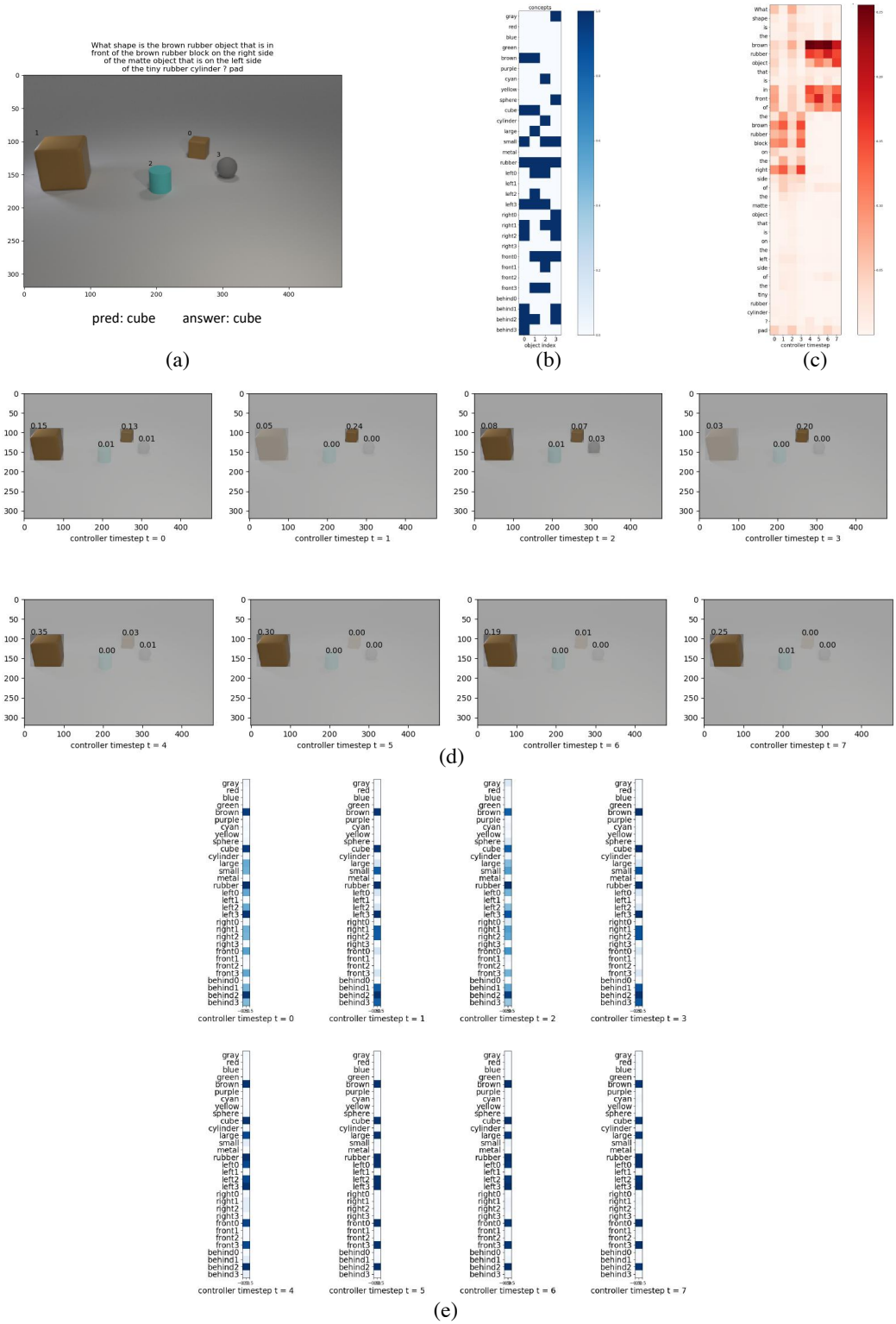


Figure 16: Visualization of reasoning steps on CLEVR dataset. (a) The question, image, prediction and ground truth answer. The index of each object is shown on the upper left of the object. (b) The induced concepts of objects and relations. (c) The stepwise attentions on question words. (d) The stepwise attentions on objects. (e) The concept vector read into the memory of the reasoning module in each step.

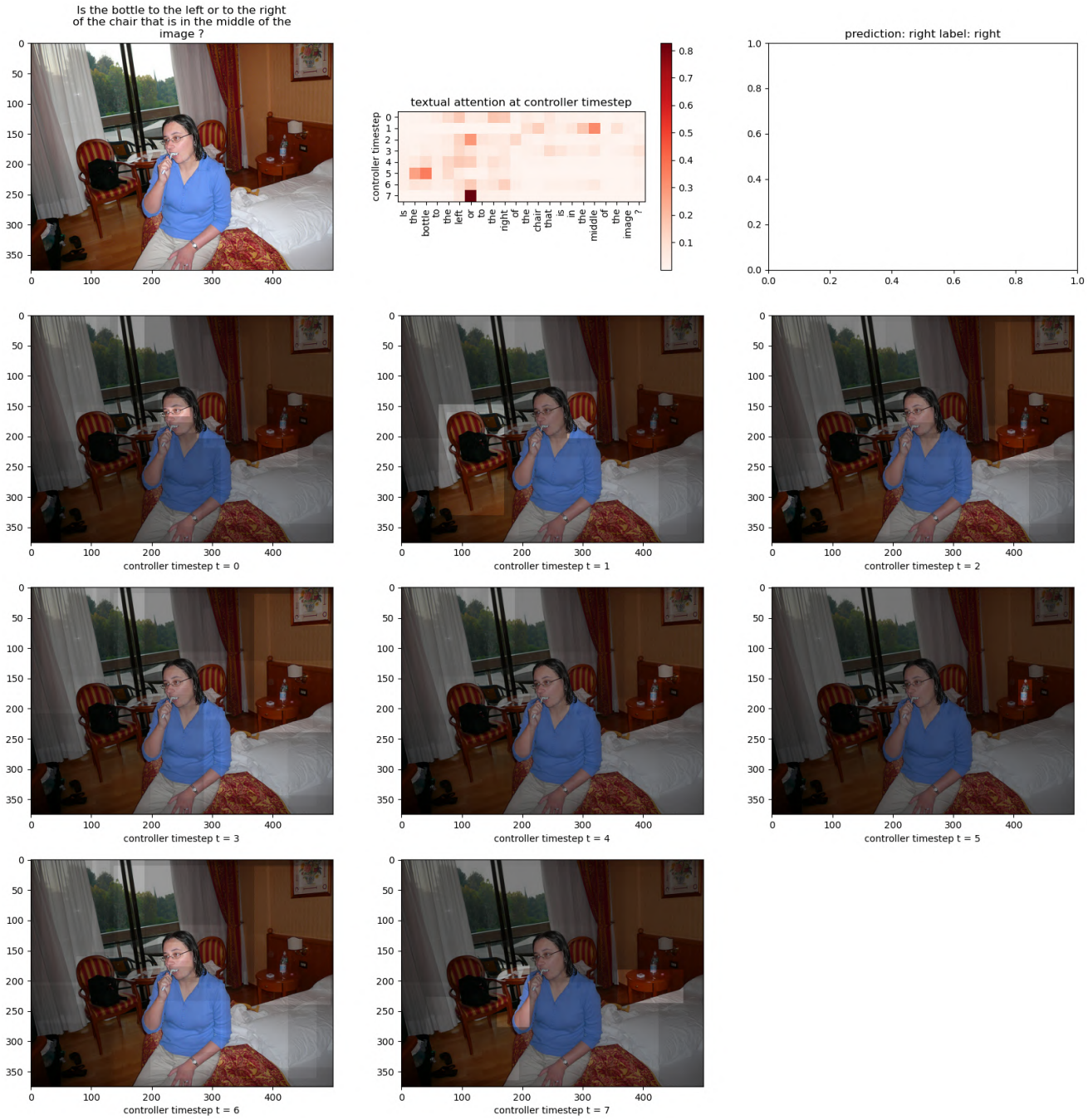


Figure 17: Visualization of reasoning steps on GQA dataset.

of order indices $O_i^{human} = (0, 1, 2, 3)$ to a_i. Then, we rank $(a_{i1}, a_{i2}, a_{i3}, a_{i4})$ with a descending order of their conditional probabilities and assign a sequence of order indices O_i^{induce} to a_i. For comparison, we further rank $(a_{i1}, a_{i2}, a_{i3}, a_{i4})$ in a descending order of cosine similarities between the GLoVe embeddings of $(a_{i1}, a_{i2}, a_{i3}, a_{i4})$ and w_i and assign a sequence of order indices $O_i^{word2vec}$ to a_i. The average ranking distance can be calculated with Eqn. (7).

$$D(O^x) = \frac{1}{|S|} \sum_{i \in S} \mathcal{K}(O_i^{human}, O_i^x), \tag{7}$$

where D represents the average ranking distance, $x \in \{induce, word2vec\}$, $\mathcal{K}$ represents the operation for calculating the normalized Kendall tau distance between two

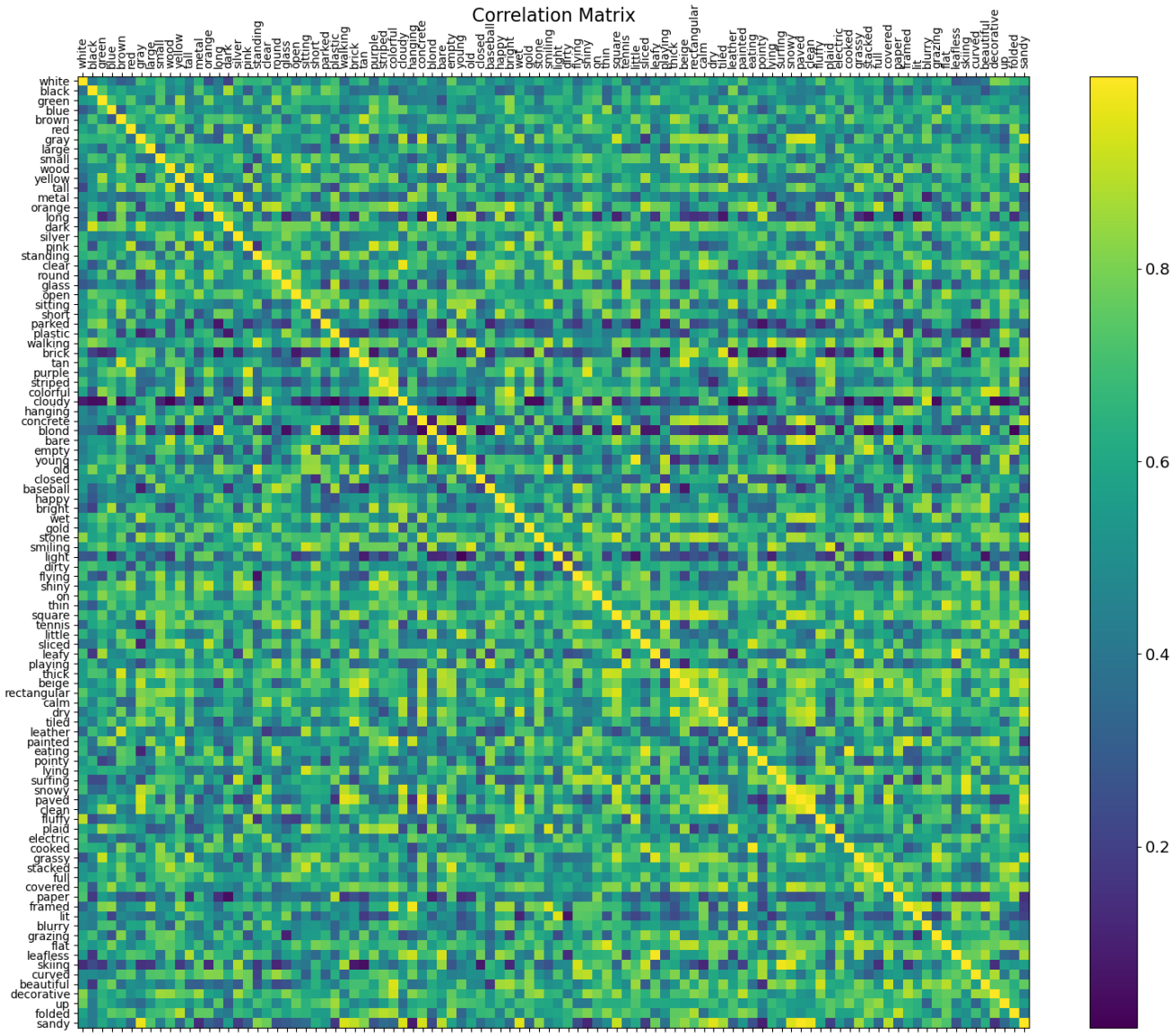


Figure 18: The extended subset of GQA concept correlations.

Table 4: The accuracy of classifying synonyms and uncorrelated words. A^{pos} represents the accuracy of classifying only synonyms. A^{neg} represents the accuracy of classifying only antonyms. †For word2vec, we tune the threshold on ground truth, while our method is used out of the box without threshold tuning (i.e., threshold set to 0.5).

Method	A^{pos}	A^{neg}	A
word2vec†	76.02%	60.71%	68.37%
induction	92.35%	63.78%	78.06%

rankings. The result in Table (5) proves that our induction

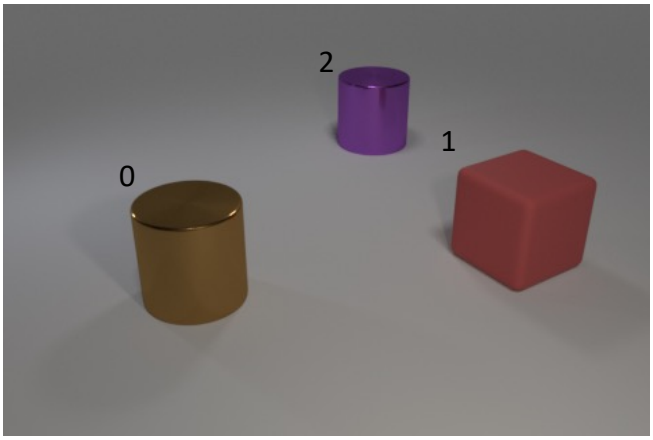
Table 5: The average ranking distance to human rankings.

$D(O^{word2vec})$	$D(O^{induce})$
0.3418	0.2585

from visual language relations encodes word proximity that is more aligned with human knowledge than the one encoded by GloVe embeddings from language-only data.

H. Error analysis

The reasoning process may reach a false answer if 1) a concept is mentioned in the question and 2) that concept



	obj0	obj1	obj2
gray	0	0	0
red	0	1	0
blue	0	0	0
green	0	0	0
brown	1	0	0
purple	0	0	1
cyan	0	0	0
yellow	0	0	0
sphere	0	0	0
cube	0	1	0
cylinder	1	0	1
large	1	1	1
small	0	0	0
metal	1	0	1
matte	0	1	0

	obj0	obj1	obj2
left0	0	0	0
left1	1	0	1
left2	1	0	0
right0	0	1	1
right1	0	0	0
right2	0	1	0
front0	0	1	0
front1	1	0	0
front2	1	1	0
behind0	0	0	1
behind1	0	0	1
behind2	0	0	0

question	answer	ground truth
There is a big brown metal cylinder; how many large matte cubes are behind it?	0	1
What is the color of the rubber cube?	red	red

Figure 19: Error analysis. The predicted unary and binary concepts corresponding to each object in the image above are shown in the tables at the middle; the digits colored in red are wrong predicted concepts. The questions, the predicted answers and the ground truth answers are shown in the table at the bottom.

is wrongly classified for the objects ought to be attended. However, the reasoning process may still reach a correct answer if either of these two conditions is not sufficed. We present two examples in Figure 19.