GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering – Supplementary Material

Drew A. Hudson
Stanford University
353 Serra Mall, Stanford, CA 94305
dorarad@cs.stanford.edu

Christopher D. Manning
Stanford University
353 Serra Mall, Stanford, CA 94305
manning@cs.stanford.edu

1. Dataset Visualizations

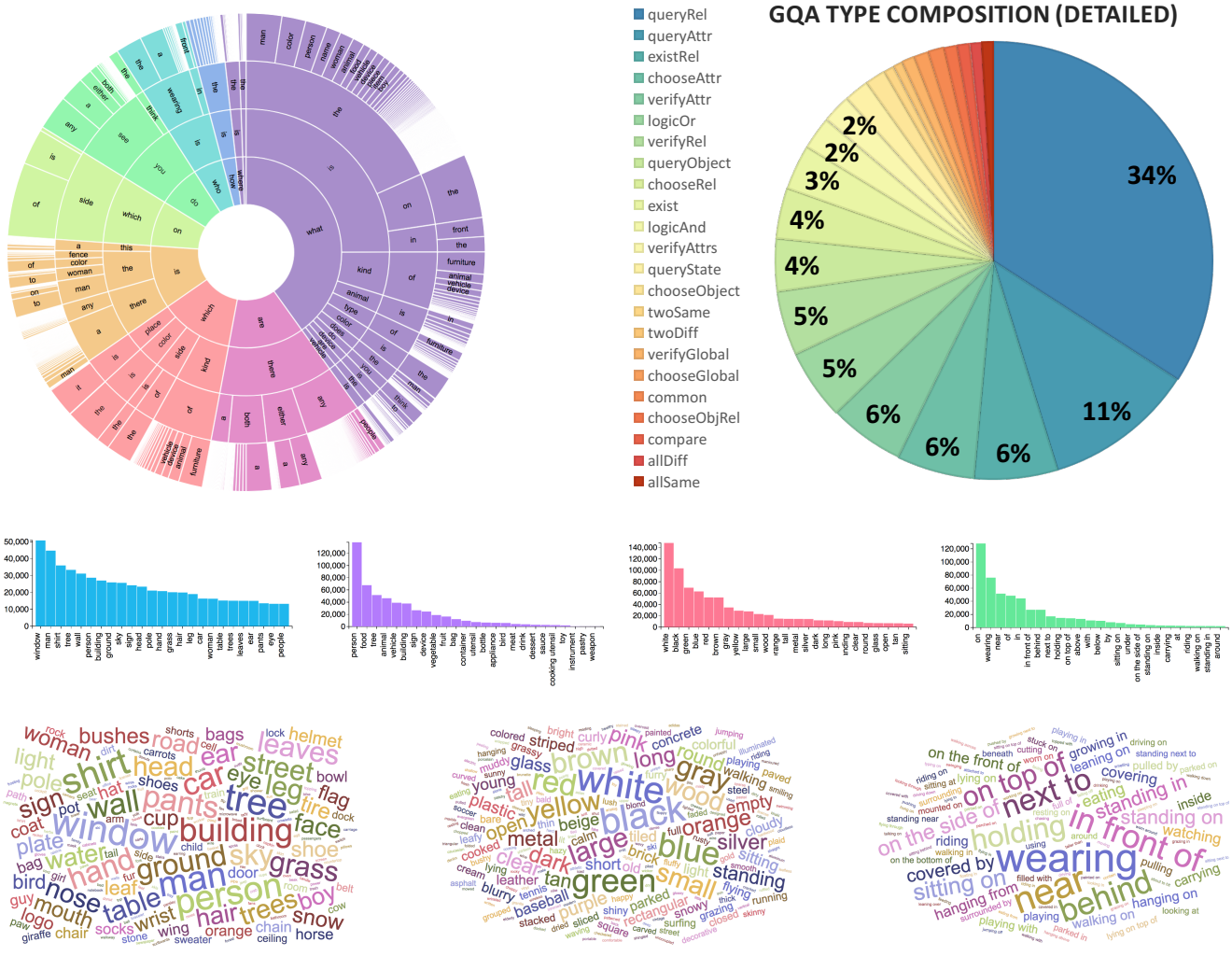


Figure 1: **Top left**: Distribution of GQA questions by first four words. The arc length is proportional to the number of questions containing that prefix. **Top right**: question type distribution; please refer to table 1 for details about each type. **Middle row**: Number of occurrences of the most frequent objects, categories, attributes and relations (excluding left/right (500k occurrences, inferred automatically rather than being hand-annotated as all other relations). **Third row**: Word clouds for frequent objects, attributes and relations.

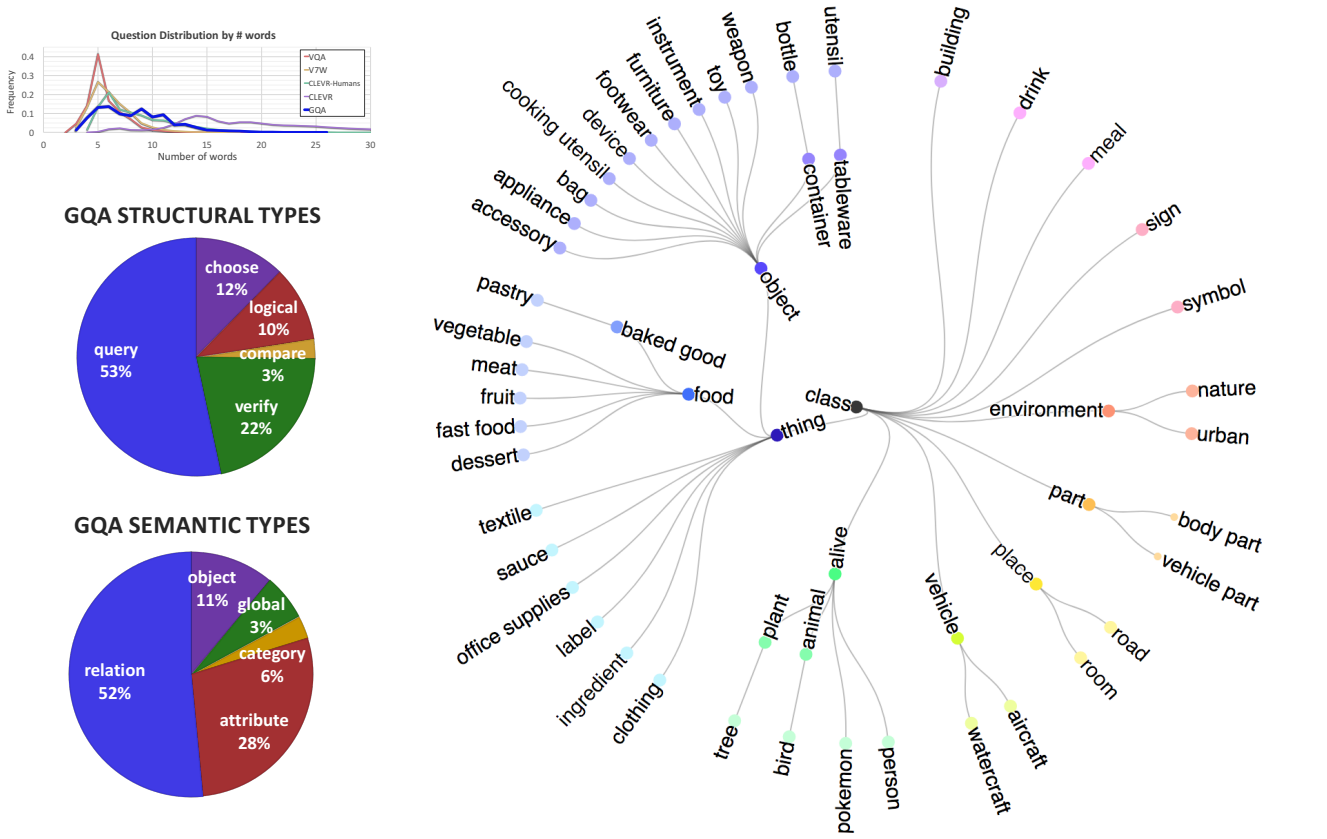


Figure 2: **Top left**: Question length distribution for VQA datasets: we can see that GQA has a diverse range of lengths compared to all other datasets except synthetic CLEVR. **Bottom left**: GQA Question structural and semantic type distributions. **Right**: The object class hierarchy we have created as part of the dataset construction process.

Type	Open/Binary	Semantic	Structural	Form	Example
queryGlobal	open	query	global	select: `scene`/query: `type`	How is the weather in the image?
verifyGlobal	binary	verify	global	select: `scene`/verify `type`: `attr`	Is it cloudy today?
chooseGlobal	open	query	global	select: `scene`/choose `type`: `a|b`	Is it sunny or cloudy?
queryAttr	open	query	attribute	select: `obj/.../query: type`	What color is the apple?
verifyAttr	binary	verify	attribute	select: `obj/.../verify type: attr`	Is the apple red?
verifyAttrs	binary	logical	attribute	select: `obj/.../verify t1: a1/verify t2: a2/and`	Is the apple red and shiny?
chooseAttr	open	choose	attribute	select: `obj/.../choose type: a|b`	Is the apple green or red?
exist	binary	verify	object	select: `obj/.../exist`	Is there an apple in the picture?
existRel	binary	verify	relation	select: `subj/.../relate (rel): obj/exist`	Is there an apple on the black table?
logicOr	binary	logical	object	select: `obj1/.../exist/select: obj2/.../exist/or`	Do you see either an apple or a banana there?
logicAnd	binary	logical	obj/attr	select: `obj1/.../exist/select: obj2/.../exist/and`	Do you see both green apples and bananas there?
queryObject	open	query	category	select: `category/.../query: name`	What kind of fruit is on the table?
chooseObject	open	choose	category	select: `category/.../choose: a|b`	What kind of fruit is it, an apple or a banana?
queryRel	open	query	relation	select: `subj/.../relate (rel): obj/query: name`	What is the small girl wearing?
verifyRel	binary	verify	relation	select: `subj/.../verifyRel (rel): obj`	Is she wearing a blue dress?
chooseRel	open	choose	relation	select: `subj/.../chooseRel (r1|r2): obj`	Is the cat to the left or to the right of the flower?
chooseObjRel	open	choose	relation	select: `subj/.../relate (rel): obj/choose: a|b`	What is the boy eating, an apple or a slice of pizza?
compare	binary	compare	object	select: `obj1/.../select: obj2/.../compare type`	Who is taller, the boy or the girl?
common	open	compare	object	select: `obj1/.../select: obj2/.../common`	What is common to the shirt and the flower?
twoSame	verify	compare	object	select: `obj1/.../select: obj2/.../same`	Does the shirt and the flower have the same color?
twoDiff	verify	compare	object	select: `obj1/.../select: obj2/.../different`	Are the table and the chair made of different materials?
allSame	verify	compare	object	select: `allObjs/same`	Are all the people there the same gender?
allDiff	verify	compare	object	select: `allObjs/different`	Are the animals in the image of different types?

Table 1: Functions Catalog for all the GQA question types. For each question we mention its structural and semantic types (refer to table 1 for further details), a functional program template and a typical example of a generated question.

GQA
1. What is the **woman** to the right of the **boat** holding? umbrella
2. Are there **men** to the left of the **person** that is holding the **umbrella**? no
3. What color is the **umbrella** the **woman** is holding? purple

VQA
1. Why is the person using an umbrella?
2. Is the picture edited?
3. What's the color of the umbrella?

GQA
1. Is that a **giraffe** or an **elephant**? giraffe
2. Who is feeding the **giraffe** behind the **man**? lady
3. Is there any **fence** near the **animal** behind the **man**? yes
4. On which side of the image is the **man**? right
5. Is the **giraffe** behind the **man**? yes

VQA
1. What animal is the lady feeding?
2. Is it raining?
3. Is the man wearing sunglasses?

GQA
1. Is the **person**'s **hair** brown and long? yes
2. What **appliance** is to the left of the **man**? refrigerator
3. Is the **man** to the left or to the right of a **refrigerator**? right
4. Who is in front of the **appliance** on the left? man
5. Is there a **necktie** in the picture that is not red? yes
6. What is the **person** in front of the **refrigerator** wearing? suit
7. What is hanging on the **wall**? picture
8. Does the **vest** have a different color than the **tie**? no
9. What is the color of the **shirt**? white
10. Is the color of the **vest** different than the **shirt**? yes

VQA
1. Does this man need a haircut?
2. What color is the guys tie?
3. What is different about the man's suit that shows this is for a special occasion?

GQA
1. Who wears the **gloves**? player
2. Are there any **horses** to the left of the **man**? no
3. Is the **man** to the right of the **player** that wears gloves? no
4. Is there a **bag** in the picture? no
5. Do the **hat** and the **plate** have different colors? yes

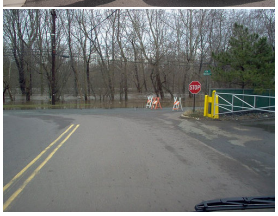
VQA
1. What is the man holding?
2. Where are the people playing?
3. Is the player safe?
4. What is the sport being played?

GQA
1. What is the **person** doing? playing
2. Is the **entertainment center** at the bottom or at the top? bottom
3. Is the **entertainment center** wooden and small? yes
4. Are the pants blue? no
5. Do you think the **controller** is red? no

VQA
1. What colors are the walls?
2. What game is the man playing?
3. Why do they stand to play?

GQA
1. Are there any **coats**? yes
2. Do you see a red **coat** in the image? no
3. Is the **person** that is to the left of the **man** exiting a **truck**? no
4. Which place is this? road

VQA
1. Where is the bus driver?
2. Why is the man in front of the bus?
3. What numbers are repeated in the bus number?



GQA
1. What is in front of the green **fence**? gate
2. Of which color is the **gate**? silver
3. Where is this? street
4. What color is the **fence** behind the **gate**? green
5. Is the **fence** behind the **gate** both brown and metallic? no

VQA
1. What are the yellow lines called?
2. Why don't the trees have leaves?
3. Where is the stop sign?

Figure 3: Examples of questions from GQA and VQA, for the same images. As the examples demonstrate, GQA questions tend to involve more elements from the image compared to VQA questions, and are longer and more compositional as well. Conversely, VQA questions tend to be a bit more ambiguous and subjective, at times with no clear and conclusive answer. Finally, we can see that GQA provides more questions for each image and thus covers it more thoroughly than VQA.

2. Dataset Balancing

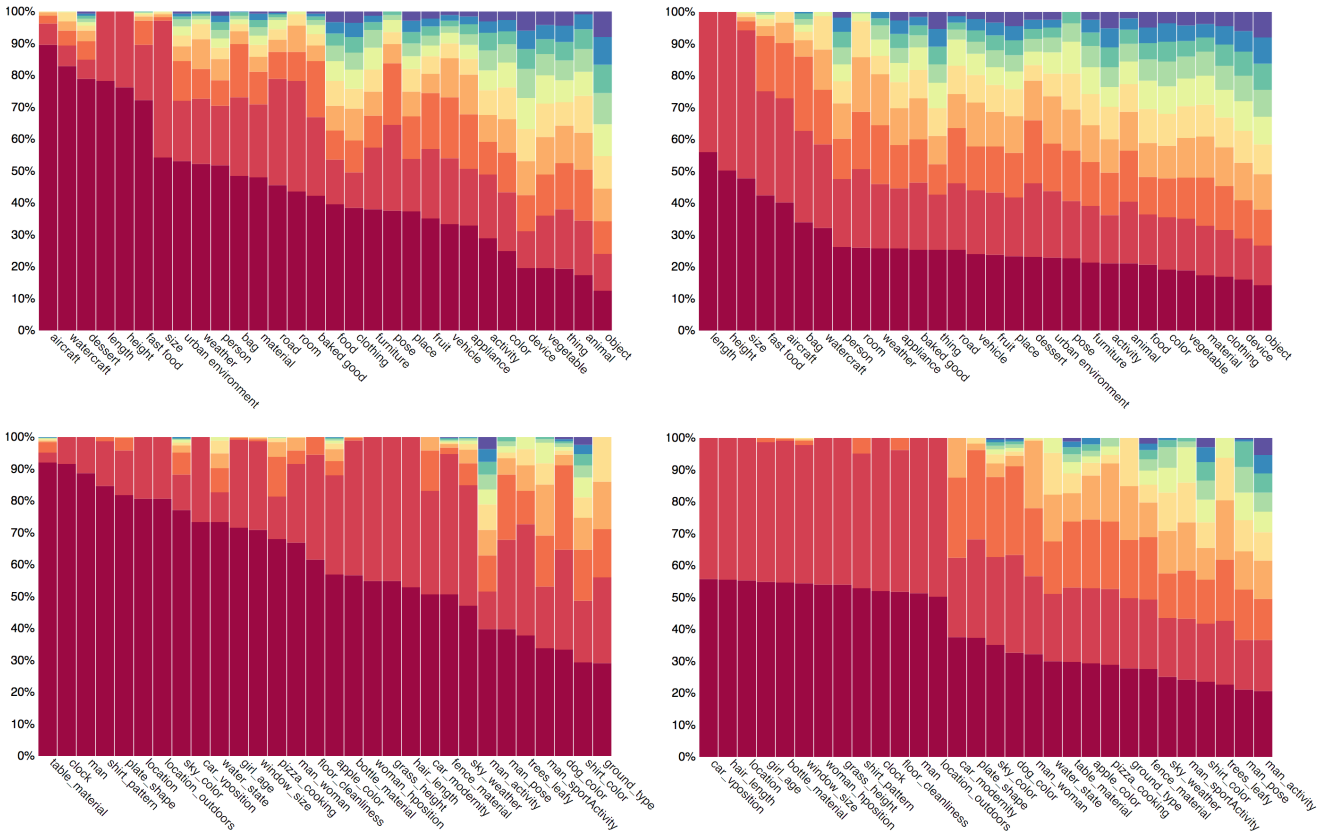


Figure 4: Impact of the dataset balancing on the conditional answer distribution: The left side shows the distribution before any balancing. We show the top 10 answers for a selection of question groups, where the column height corresponds to the relative frequency of each answer. The top row shows global question groups such as color questions, questions about animals, etc. while the bottom row shows local ones *e.g. apple-color*, *table-material* etc. (section 3.3, main paper). As we can see, these initial distributions are heavily biased. The right side shows the distributions after balancing, more uniform and with heavier tails, while intentionally retaining the original real-world tendencies up to a tunable degree.

As discussed in section 3.4 (main paper), given the original 22M auto-generated questions, we have performed answer-distribution balancing, similarities reduction and type-based sampling, producing a 1.7M questions balanced dataset. The balancing is performed in an iterative manner: as explained in section 3.3, for each question group (e.g., color questions), we iterate over the answer distribution, from the most to least frequent answers: (a_i, c_i) when a_i is the answer and c_i is its count. In each iteration i, we downsample the head distribution $(a_j, j \leq i)$ such that the ratio between the head and its complementary tail $\frac{\sum_{j \leq i} c_j}{1-\sum_{j \leq i} c_j}$ will be bounded by b. While doing so, we also make sure to set minimum and maximum bounds on the frequency ratio $\frac{c_{i+1}}{c_i}$ of each pair of consequent answers a_i, a_{i+1}. The results of this process is shown in figure 4, demonstrating how the distribution is "pushed" away from the head and spreads over the tail, while intentionally maintaining the original real-world tendencies presented in the data, to retain its authenticity.

3. Baselines Implementation Details

In section 4.2 (main paper), we perform experiments over multiple baselines and state-of-the-art models. All CNN models use spatial features pre-trained on ImageNet [3], whereas the state-of-the-art approaches bottomUp [2] and MAC [5] are based on object-based features produced by the Faster R-CNN detector [11]. All models use GloVe word embeddings of dimension 300 [10]. To allow a fair comparison, all the models use the same LSTM, CNN and classifier components, and so the only difference between the models stem from their core architectural design.

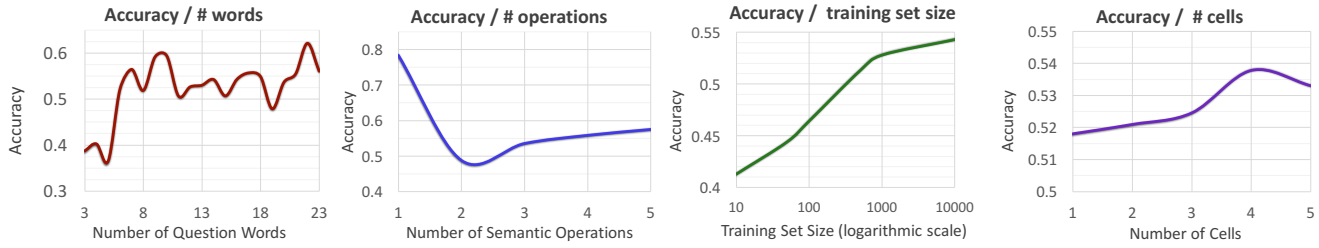


Figure 5: From left to right: **(1)** Accuracy as a function of textual question length – the number of words in the question. **(2)** Accuracy as a function of semantic question length – the number of operations in its functional program. **(3)** Performance as a function of the subset size used for training, ranging from 10K to 10M. **(4)** Accuracy for different lengths of MAC networks, suggesting that indeed GQA questions are compositional.

We used a sigmoid-based classifier and trained all models using Adam [7] for 15 epochs, each taking about an hour to complete. For MAC [5], we use the authors' code available online, with 4 cells. For BottomUp [2], since the original implementation is unfortunately not publicly available, we re-implemented the model, carefully following details presented in [2, 12]. To ensure the correctness of our implementation, we have tested the model on the standard VQA dataset, achieving 67%, which matches the original scores reported by Anderson *et al*. [2].

4. Further Diagnosis

Following section 4.2 (main paper), and in order to get more insight into models' behaviors and tendencies, we perform further analysis of the top-scoring model for the GQA dataset, MAC [5]. The MAC network is a recurrent attention network that reasons in multiple concurrent steps over both the question and the image, and is thus geared towards compositional reasoning as well as rich scenes with several regions of relevance.

We assess the model along multiple axes of variation, including question length, both textually, *i.e*. number of words, and semantically, *i.e*. number of reasoning operations required to answer it, where an operation can be *e.g*. following a relation from one object to another, attribute identification, or a logical operation such as *or*, *and* or *not*. We provide additional results for different network lengths (namely, cells number) and varying training-set sizes, all can be found in figure 5.

Interestingly, question textual length correlates positively with the model accuracy. It may be the case that longer questions reveal more cues or information that the model can exploit, potentially sidestepping direct reasoning about the image. However, question semantic length has the opposite impact as expected: 1-step questions are particularly easy for models than the compositional ones which involve more steps.

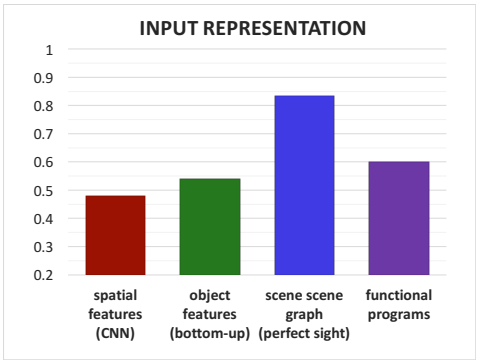


Figure 6: Performance as a function of the input representation. We encode the scenes through three different methods: spatial features produced by a standard pretrained CNN, object-based features generated by a faster R-CNN detector, and direct embedding of the scene graph semantic representation, equivalent to having perfect sight. We further experiment with both textual questions as well as their counterpart functional programs as input. We can see that the more semantically-imbued the representations get, the higher the accuracy obtained.

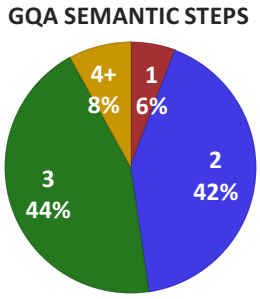


Figure 7: Distribution of GQA questions semantic length (number of computation steps to arrive at the answer). We can see that most questions require about 2-3 reasoning steps, where each step may involve tracking a relation between objects, an attribute identification or a logical operation.

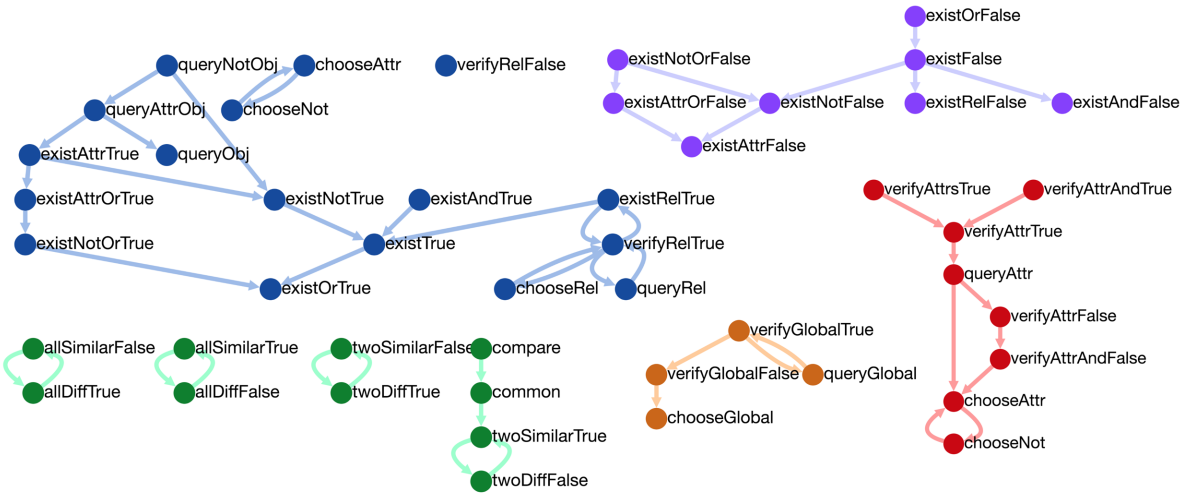


Figure 8: Entailment relations between different question types. In section 3.3 (main paper) we discuss the entailment and equivalences between questions. Since every question in the dataset has a matching logical representation of the sequence of reasoning steps, we can formally compute all the entailment and equivalence relations between different questions. Indeed, a cogent and reasonable learner should be consistent between its own answers, *e.g.* should not answer "red" to a question about the color of an object it has just identified as blue. Some more subtle relations also occur, such as those involving relations, *e.g.* if `X` is above `Y`, than `Y` is below `X`, and `X` is not below `Y`, etc. figure 8 shows all the logical relations between the various question types. Refer to table 1 for a complete catalog of the different types. Experiments show that while people excel at consistency, achieving the impressive 98.4%, deep learning models perform much worse in this task, with 69% - 82%. These results cast a doubt about the reliability of existing models and their true visual understanding skills. We therefore believe that improving their skills towards enhanced consistency and cogency is an important direction, which we hope our dataset will encourage.

We can further see that longer MAC networks with more cells are more competent in performing the GQA task, substantiating its increased compositionality. Other experiments show that increasing the training set size has significant impact on the model's performance, as found out also by Kafle *et al*. [6]. Apparently, the training set size has not reached saturation yet and so models may benefit from even larger datasets.

Finally, we have measured the impact of different input representations on the performance. We encode the visual scene with three different methods, ranging from standard pretrained CNN-based spatial features, to object-informed features obtained through faster R-CNNs detectors [11], up to even a "perfect sight" model that has access to the precise semantic scene graph through direct node and edge embeddings. As figure 5 shows, the more high-level and semantic the representation is, the better are the results.

On the question side, we explore both training on the standard textual questions as well as the semantic functional programs. MAC achieves 53.8% accuracy and 81.59% consistency on the textual questions and 59.7% and 85.85% on the programs, demonstrating the usefulness and further challenge embodied in the former. It is also more consistent Indeed, the programs consist of only a small operations vocabulary, whereas the questions use both synonyms and hundreds of possible structures, incorporating probabilistic rules to make them more natural and diverse. In particular, GQA questions have sundry subtle and challenging linguistic phenomena such as long-range dependencies, absent from the canonical programs. The textual questions thus provide us with the opportunity to engage with real, interesting and significant aspects of natural language, and consequently foster the development of models with enhanced language comprehension skills.

5. Comparison between GQA and VQA 2.0

We perform a comparison between GQA and VQA 2.0 [4] datasets, as summarized in table 2. We can see that GQA questions are longer on average, and consequently contain more verbs and prepositions than VQA (as well as more nouns and adjectives), providing further evidence for its increased compositionality. Semantically, GQA questions are significantly more compositional than VQA's, and involve a variety of reasoning skills with much higher frequency (spatial, logical, relational and comparative).

Some VQA question types are not covered by GQA, such as intention (*why*) questions (*Why is she holding an umbrella?*), subjective questions (*Does he like blue?*) or ones involving OCR (*What does the sign read?*) or external knowledge (*In which university are they studying? Which animal likes bananas?*). In contrast, the GQA dataset focuses on factual questions and multi-hop reasoning in particular, rather than covering all types. Comparing to VQA, GQA questions are by-construction more objective, unambiguous, and can be answered from the images only, potentially making this benchmark more controlled and convenient for making research progress on.

Aspect	VQA 2.0	GQA
Question length	6.2 + 1.9	7.9 + 3.1
Verbs	1.4 + 0.6	1.6 + 0.7
Nouns	1.9 + 0.9	2.5 + 1.0
Adjectives	0.6 + 0.6	0.7 + 0.7
Prepositions	0.5 + 0.5	1.0 + 1.0
Relation questions	19.5%	51.6%
Spatial questions	8%	22.4%
Logical questions	6%	19%
Comparative questions	1%	3%
OCR questions	3%	0%
Intention questions (*why*)	1.2%	0%
Where questions	2.9%	1.6%
Who questions	0.8%	6.1%
Counting questions	10.3%	0%
Short questions ($\leq$ 5 words)	60%	22%
Compositional questions (semantically)	3%	52%

Table 2: A comparison between GQA and VQA 2.0. GQA questions are longer on average, more compositional and require more reasoning: spatial, logical, relational, comparative. VQA questions cover additional types such as OCR, intention (*why*) and counting (which we may include in an updated version of GQA). Question types are not mutually exclusive.

6. Scene Graph Normalization

Our starting point in creating the GQA dataset is the Visual Genome Scene Graph annotations [8] that cover 113k images from COCO [9] and Flickr [13]. The scene graph serves as a formalized representation of the image: each node denotes an **object**, a visual entity within the image, like a person, an apple, grass or clouds. It is linked to a bounding box specifying its position and size, and is marked up with about 1-3 **attributes**, properties of the object: e.g., its color, shape, material or activity. The objects are connected by **relation** edges, representing actions (verbs), spatial relations (prepositions), and comparatives.

The scene graphs are annotated with free-form natural language. Our first goal is thus to convert the annotations into a clear and unambiguous semantic ontology. We begin by cleaning up the graph's vocabulary, removing stop words, fixing typos, consolidating synonyms and filtering rare or amorphous concepts. We then classify the vocabulary into predefined categories (e.g., *animals* and *fruits* for objects; *colors* and *materials* for attributes), using word embedding distances to get preliminary annotations, which are then followed by manual curation. This results in a class hierarchy over the scene graph's vocabulary, which we further augment with various semantic and linguistic features like part of speech, voice, plurality and synonyms – information that will be used to create grammatically correct questions in further steps. Our final ontology contains 1740 objects, 620 attributes and 330 relations, grouped into a hierarchy that consists of 60 different categories and subcategories. Visualization of the ontology can be found in figure 2.

At the next step, we prune graph edges that sound unnatural or are otherwise inadequate to be incorporated within the questions to be generated, such as *(woman, in, shirt)*, *(tail, attached to, giraffe)*, or *(hand, hugging, bear)*. We filter these triplets using a combination of category-based rules, n-gram frequencies [1], dataset co-occurrence statistics, and manual curation.

In order to generate correct and unambiguous questions, some cases require us to validate the uniqueness or absence of an object. Visual Genome, while meant to be as exhaustive as possible, cannot guarantee full coverage (as it may be practically infeasible). Hence, in those cases we use object detectors [11], trained on visual genome with a low detection threshold, to conservatively confirm the object absence or uniqueness. Here, object uniqueness can be validated by confidently denying the existence of other same-class objects within the image.

Next, we augment the graph with absolute and relative positional information: objects appearing within the image margins, are annotated accordingly. Object pairs for which we can safely determine positional relations (e.g., one is to the left of the other), are annotated as well. We also annotate object pairs if they share the same color, material or shape. Finally, we enrich the graph with global information about the image location or weather, if these can be directly inferred from the objects it contains.

By the end of this stage, the resulting scene graphs have clean, unified, rich and unambiguous semantics for both the nodes and the edges.

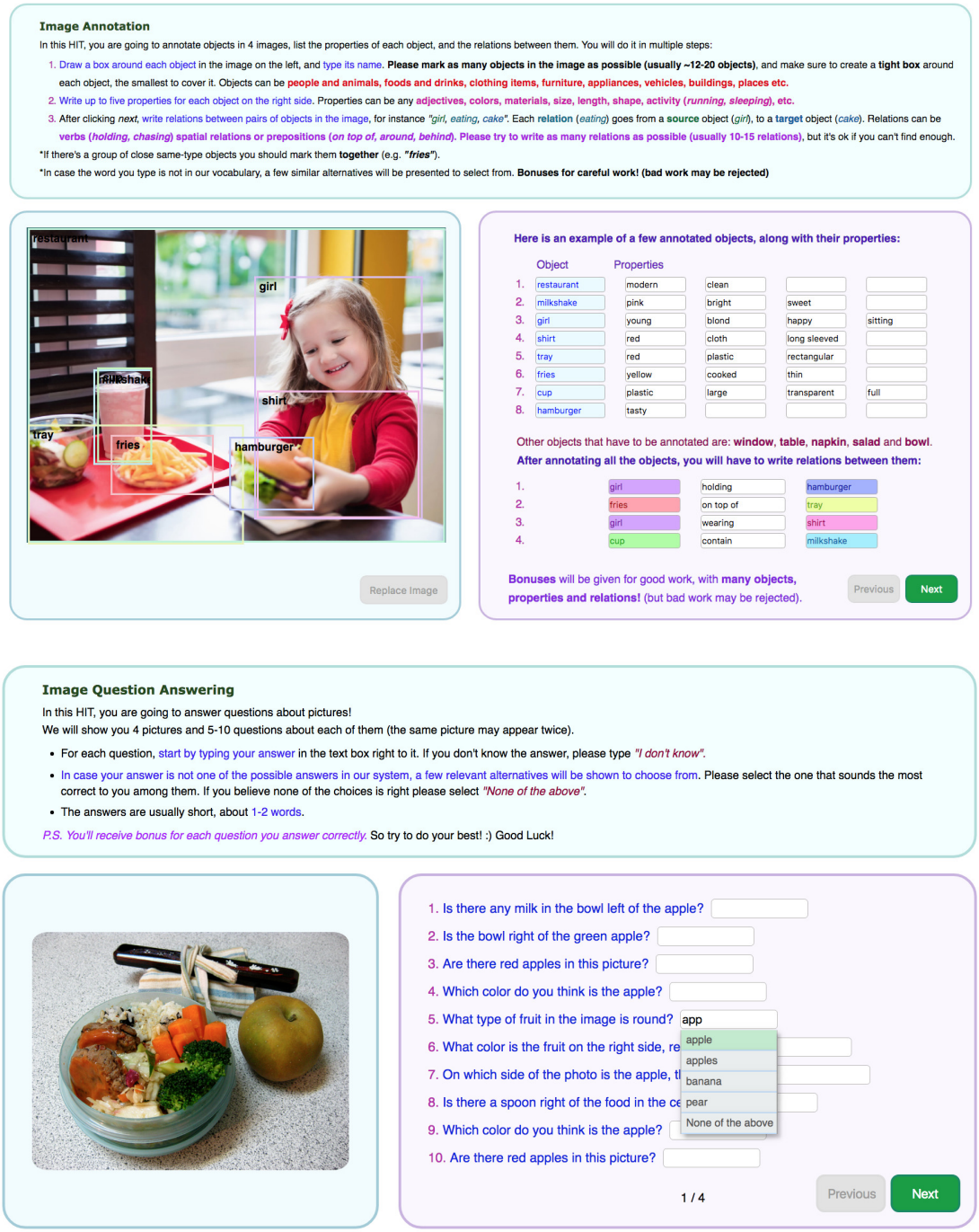


Figure 9: The interfaces used for human experiments on Amazon Mechanical Turk. **Top**: Each HIT displays several images and asks turkers to list objects and annotate their corresponding bounding boxes. In addition, the turkers are requested to specify attributes and relations between the objects. An option to switch between images is also given to allow the turkers to choose rich enough images to work on. **Bottom**: Each HIT displays multiple questions and requires the turkers to respond. Since there is a closed set of possible answers (from a vocabulary of 1878 possibilities), and in order to allow a fair comparison between human and models' performance, we give turkers the option to respond in unconstrained free-form language, but also suggest them multiple answers from our vocabulary that are the most similar to theirs (using word embedding distances). However, turkers are not limited to choose from the suggestions in case they believe none of the proposed answers is correct.

References

[1] Google books ngram corpus available at http://books.google.com/ngrams/. 7

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017. 4, 5

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 4

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6325–6334, 2017. 7

[5] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *International Conference for Representation Learning (ICLR)*, 2018. 4, 5

[6] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. 6

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 7

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[10] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4, 6, 7

[12] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017. 5

[13] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 7