QLEVR: A Diagnostic Dataset for Quantificational Language and Elementary Visual Reasoning

Zechen Li
Northeastern University, US
li.zec@northeastern.edu

Anders Søgaard
University of Copenhagen, Denmark
soegaard@di.ku.dk

Abstract

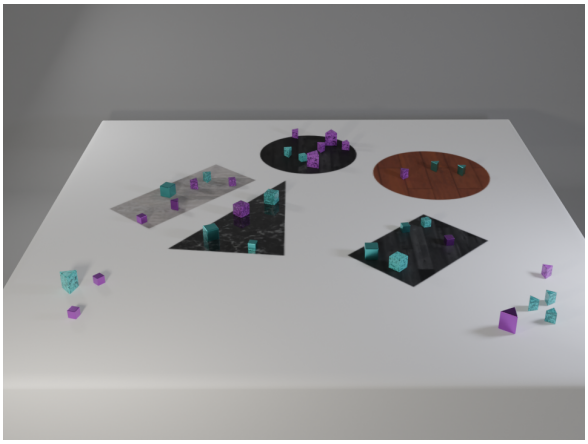
Synthetic datasets have successfully been used to probe visual question-answering datasets for their reasoning abilities. CLEVR (Johnson et al., 2017), for example, tests a range of visual reasoning abilities. The questions in CLEVR focus on comparisons of shapes, colors, and sizes, numerical reasoning, and existence claims. This paper introduces a minimally biased, diagnostic visual question-answering dataset, QLEVR, that *goes beyond existential and numerical* quantification and focus on more complex quantifiers and their combinations, e.g., asking whether there are *more than two red balls that are smaller than at least three blue balls* in an image. We describe how the dataset was created and present a first evaluation of state-of-the-art visual question-answering models, showing that QLEVR presents a formidable challenge to our current models. Code and Dataset are available at https://github.com/zechenli03/QLEVR

1 Introduction

Visual question answering is at the locus of computer vision and natural language processing, and its objective is developing computer vision systems that can answer arbitrary natural language questions about images (Lu et al., 2016; Schwartz et al., 2017; Ramakrishnan et al., 2018; Gat et al., 2020). This is useful across a range of applications, including medical image analysis, accessibility for visually impaired, video surveillance, art and advertisement (Barra et al., 2021).

The complexity of visual question answering naturally depends on the complexity of the images and the complexity of the natural language questions. The task reduces to object recognition for very simple questions of the form:

(1) Is there a triangle in this image?



Question: Are all the cyan metallic triangular prisms on the brown plane?
Answer: True

Question: On the non-white planes on the left rear side of the black wood rectangular plane, all the cyan metallic cubes but at least 2 are larger than at most 7 cubes; is it right?
Answer: False

Figure 1: A sample image and questions from QLEVR. Tasks involve attribute recognition, counting, comparing numbers, spatial relationships, and understanding of quantifiers.

Object recognition can of course be a very complex task on its own, depending on the types of objects, the number of possible objects to be recognized, the amount of supervision for inducing a good model, general image quality, etc. However, more complex queries such as (2) make visual question answering much harder:

(2) Is there a triangle inside a circle in this image?

Answering such a question in the presence of an image requires a computer vision system that not only recognizes objects, but also relations between them. CLEVR (Johnson et al., 2017) probes computer vision systems's ability to answer even more complex queries, such as, for instance:

(3) Is there a cyan cube to the right of the yellow sphere?

Findings of the Association for Computational Linguistics: NAACL 2022, pages 980 - 996

Question (3) involves reasoning about the relation between two objects, as well as the compositional semantics of color adjectives. In addition to shapes and colors, CLEVR also includes questions about sizes and quantities.

In this paper, we present a novel visual question-answering dataset that goes beyond CLEVR in focusing specifically on *quantificational language*, e.g.:

(4) Are most of the cyan cubes to the right of the yellow sphere?

Given the complexity of quantificational language, the rich typology of expressions of quantification across different languages, and the interest from philosophy, it is perhaps surprising that quantificational language has received relatively little attention in the NLP community (see §2), but we believe it is a crucial step in pushing the research horizons in (visual) question-answering.

Contributions Based on a comprehensive typology of English quantifiers, we build a dataset of 100,000 synthetic images and 999,446 unique questions to these images. This is roughly the same size as or a little bigger than CLEVR (Johnson et al., 2017). Our questions are on average longer than previous datasets. We evaluate three baselines from Johnson et al. (2017), a text-only baseline based on BERT (Devlin et al., 2019), and MAC (Hudson and Manning, 2018) on QLEVR and analyze performance across quantifier types.

2 Related Work

Visual Question Answering Challenge Datasets Several synthetic challenge datasets for visual question answering exist: Andreas et al. (2016) presents SHAPES, a predecessor to CLEVR and QLEVR, relying also on synthetic constellations of colored geometric shapes and template-driven question generation. Pezzelle and Fernández (2019) create a similar dataset to probe visual question answering models for knowledge of adjectival semantics. A portion of the visual question answering dataset (Agrawal et al., 2017) contains synthetic cartoon imagery. Sampat et al. (2021) present an extension of CLEVR that probes for hypothetical reasoning of the form: *If someone removed three triangles from this image, how many would be left?* Malinowski and Fritz (2014) combined natural images with synthetic, template-driven question generation. Finally, Parfenova et al. (2021) recently created a dataset of three-image scenes to probe two-step reasoning.

Synthetic visual question answering datasets have several advantages over ones based on real images and questions that tend to suffer from selection biases (Liu et al., 2021), but of course they are limited in what can be induced from them. They are therefore mostly useful for probing the limitations of visual question answering architectures and off-the-shelf models. Showing results only on synthetic data is often seen as a weakness in the literature (Hassantabar, 2018), but synthetic data is useful for diagnosing the errors of visual question answering systems, in our case highlighting the challenges posed by quantifiers.

Quantifiers Quantifiers have been largely ignored in the NLP community. Question-answering datasets have been developed for numerical reasoning in English (Dua et al., 2019), and some have identified quantifier words as important sources of errors for textual entailment systems (Joshi et al., 2020). Fang and Lou (2021) recently focused on the two quantifier words *part* and *whole* in an error analysis for named entity recognition.

3 QLEVR

We design a challenge dataset called QLEVR (for Quantificational Language and Elementary Visual Reasoning) that requires more complex reasoning than previous visual question-answering datasets. QLEVR is designed to probe the visual reasoning capabilities of visual question-answering systems with respect to quantificational language, including detecting members of sets, quantifying sets, and reasoning about the relationships between sets. To this end, we automatically construct *scene graphs* (Johnson et al., 2015) and use these to generate synthetic images with ground-truth locations, attributes, and relationships for planes and objects. Each scene graph can be queried in a number of way, and we design query templates to render natural language questions involving complex reasoning about sets of such planes and objects. We describe each of these steps in detail:

Image Generation All images in QLEVR are images of objects organized in a particular way on a desk-like surface. Figure 2 shows how the images are generated. We construct a scene graph for a two-dimensional image containing areas and ob-

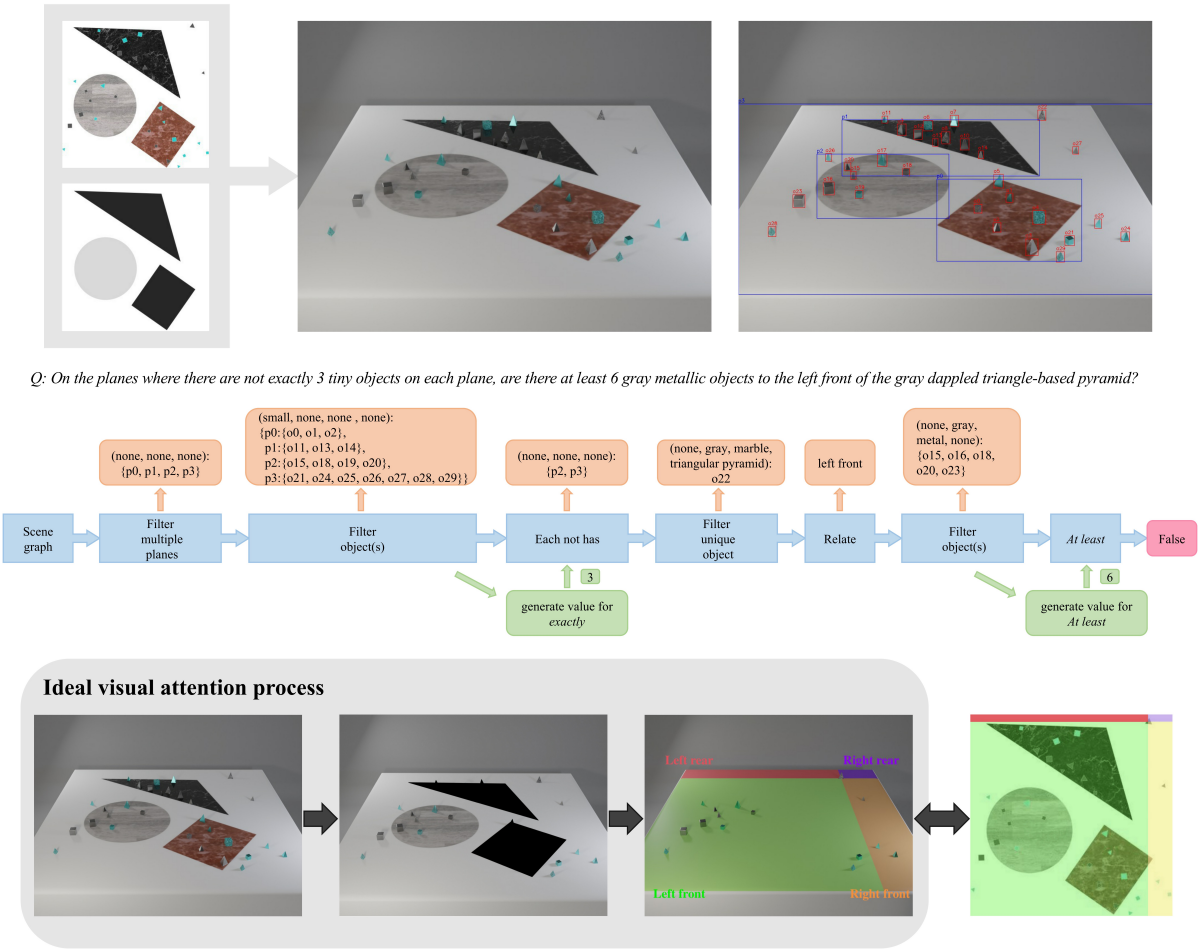


Figure 2: An overview of our dataset. **Top:** Image generation process and bounding box information. The top two-dimensional image records the scene graph, and the bottom gray-scale color map records roughness of each plane. **Center:** Examples of questions and their associated operators. **Bottom:** Ideal visual attention as the operator proceeds.

jects of different sizes and shapes. Scene graphs determine the ground-truth locations, bounding boxes, attributes and relationships for the planes and objects in the form of a graph or tree structure. Nodes are planes or objects annotated with attributes, each of which is connected to its spatially related nodes.

Each image contains one to five areas or geometric planes. These can be either triangular, rectangular or circular. The rest of the desk area we refer to as the white non-geometric plane. Geometric planes come in two materials (marble and wood), three colors (black, gray, and brown), and random sizes.

Each geometric plane contains one to ten (1–10) objects, with different sizes and shapes, and the non-geometric plane contains one to twelve (1–12) objects, with different sizes and shapes. Object come in seven shapes (cone, cube, cylinder, pentahedron, sphere, triangular prism, and tetrahedron), two absolute sizes (small and large), five materials (metal, rubber, leather, marble, and wood), and eight colors (blue, brown, cyan, gray, green, purple, red and yellow). The spatial relationships between planes and objects include *front*, *back*, *left* and *right*, as well as *right front*, *right rear*, *left front* and *left rear*.

We render three-dimensional images of the scene graphs with Blender (Community, 2018). Light settings and three preset camera positions were chosen at random, after validating that all objects were at least partially visible. Since the depth of the scene can affect the judgment of the spatial relationship in the three-dimensional image, the desk boundary is always visible as a reference for determining the depth of the scene. Minimum distances between objects and planes were kept to reduce the ambiguity of spatial relationships. See Appendix B for more details.

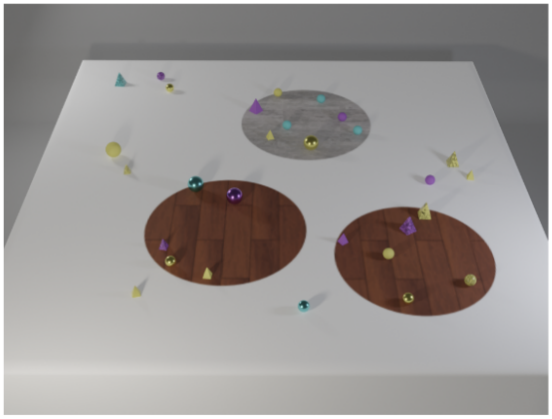
$all_P(A, B) \Leftrightarrow A \subseteq B$
$some_P(A, B) \Leftrightarrow A \cap B \neq \emptyset$
$no_P(A, B) \Leftrightarrow A \cap B = \emptyset$
$some\ but\ not\ all_P(A, B) \Leftrightarrow A \cap B \neq \emptyset \neq A - B$
$most_P(A, B) \Leftrightarrow |A \cap B| > |A - B|$
$more_P(A, B) \Leftrightarrow |A| > |B|$
$fewer_P(A, B) \Leftrightarrow |A| < |B|$
$equal_P(A, B) \Leftrightarrow |A| = |B|$
$exactly\ \mathrm{n}_P(A, B) \Leftrightarrow |A| = \mathrm{n}\ \ \&\ \ A \subseteq B$
$between\ \mathrm{n}_1\ and\ \mathrm{n}_2{}_P(A, B) \Leftrightarrow \mathrm{n}_1 \leq |A \cap B| \leq \mathrm{n}_2$
$at\ most\ \mathrm{n}_P(A, B) \Leftrightarrow |A \cap B| \leq \mathrm{n}$
$more\ than\ \mathrm{n}_P(A, B) \Leftrightarrow |A \cap B| > \mathrm{n}$
$all\ but\ at\ least\ \mathrm{n}_P(A, B) \Leftrightarrow |A - B| \geq \mathrm{n}$
$at\ least\ \frac{\mathrm{n}}{\mathrm{d}}\ of\ the_P(A, B) \Leftrightarrow \frac{|A \cap B|}{|A|} \geq \frac{\mathrm{n}}{\mathrm{d}}$
$fewer\ than\ \frac{\mathrm{n}}{\mathrm{d}}\ of\ the_P(A, B) \Leftrightarrow \frac{|A \cap B|}{|A|} < \frac{\mathrm{n}}{\mathrm{d}}$
$no\ objects\ except\ C_P(A, B) \Leftrightarrow A \cap B = \{c\}$
$every\ object\ except\ C_P(A, B) \Leftrightarrow A - B = \{c\}$

Table 1: Quantifiers included in QLEVR. P denotes the set of all the objects on the target plane(s). A or B denotes a subset of P with the same attributes. $|A|$ is the cardinality of the set A.

Question Generation Quantifiers are often said to be among the most important and complex constructs of natural languages (Hintikka, 1977; Barwise and Cooper, 1981). As pointed out by by Bernardi and Pezzelle (2021), visual question-answering models need to master a wide range of linguistic phenomena, including negation, entailment, mutual exclusivity and so on. We add (generalized) quantifiers to this list and design a dataset to probe the ability of visual question-answering systems to handle quantifiers in combination with other linguistic phenomena. See Table 1 for the quantifiers included in QLEVR.

See Figure 2 for how questions are formed from scene graphs. In brief, we think of the scene graph as a model and evaluate various combinations of logical operators, including quantifiers, on the scene graph, i.e., performing a model checking (Clarke et al., 2009) procedure.

We introduce the notion of a *question family*, defined by a set of operators and a scene graph. Each question family is associated with 2–6 text templates and a set of synonyms (for shapes, colors, materials, and spatial relationships). The templates were written by hand. Each question template can thus generate multiple questions. For example, the



Question: On the geometric plane whose color is different from that of other planes, all yellow spheres but 1 are smaller than fewer than 1 or more than 3 spheres; is it right?
Answer: True
Quantifiers: exactly N (all but N ¬), between (¬ between)

Figure 3: Example image-question pair

template

(5) Are there exactly `<OC> <Z> <C> <M> <S><os>` on the `<PC> <PM> <PS>` plane`<ps>`?

where upper-cased variables refer to words, and lower-cased variables to suffixes, can generate the question

(6) Are there exactly 2 small red rubber objects on the black wooden triangular plane?

We construct a total of 671 different templates, which are randomly constructed from 11 plane templates and 61 object templates. Our questions involve attribute recognition, counting, comparing numbers or attributes, spatial relationships, and understanding of quantifiers. Figure 2 shows the operators built in the given question family, such as *filter*, *relate*, and *at least*.

Note that many (generalized) quantifiers are related by entailment. The question

(7) Are all the red cubes on the marble planes?

is, assuming an image with red cubes, semantically equivalent to

(8) Are no red cubes not on the marble planes?

The semantics of combinations of quantifiers can be derived using *squares of opposition* (Westerståhl, 2012). We exploit these entailment relations in creating QLEVR.

Split	Images	Questions	Unique questions	Overlap with train	Overlap with val
Train	70,000	700,000	699,498	-	-
Val	15,000	150,000	149,968	199 | 148	-
Test	15,000	150,000	149,980	194 | 145	49 | 39
Total	100,000	1,000,000	999,446	-	-

Table 2: Statistics for our dataset. In each *Overlap* column, the number on the right represents the number of overlapping questions with the same answer.

Some combinations of key values may generate unreasonable questions. We therefore define restrictions for each question family to avoid the generation of *pragmatically odd*, *ill-posed* or *trivial* questions. For example, the phrase *on the marble plane where there are at least 5 red objects* would be pragmatically odd if there was only one marble plane in the scene. The sentence

(9) On the marble plane, do between 2 and 4 cubes have the same size as most of the cylinders?

is ill-posed if there are no cubes on the marble plane. Finally, questions like *Are there more red cubes than cubes?* are trivial, because they can be answered in the absence of the image. The assertion is always true. The opposite would, for example, be true of

(10) On the plane with 8 balls, are there exactly 3 balls?

We present many examples of images and questions in the Appendix, but see also Figure 3 for a complex question with embedded quantifiers.

Dataset Characteristics QLEVR has 1,000,000 questions for 100,000 images, with each image having 10 questions generated from different question families. The dataset is balanced, preventing answering in the absence of the images. In addition, the answer distribution across question families is constrained by acceptance-rejection sampling. The data is randomly split, with 70% for training data, 15% for validation and 15% for heldout evaluation data (the test set). As shown in figure 4, QLEVR includes 27 different quantifiers. Questions contain 1–4 quantifiers. Table 2 shows the diversity and complexity of the QLEVR questions. Almost all the questions are unique. Very few questions appear in several splits, and always in conjunction with new scene graphs.

4 Experiments

In this section, we evaluate the performance of baselines and near-state-of-the-art models on the QLEVR dataset and perform detailed error analysis. We ran each each method three times with different random seeds and report the test set performance for the model that achieved the best performance on the validation data.

4.1 Models

We first present three purely text-based models, Q-type (Antol et al., 2015), LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2019), to evaluate the level of visual reasoning needed for QLEVR. If these perform at random (0.5), we have successfully constructed a dataset in which questions cannot be answered in the absence of images. It is important to include text-only models as baselines in visual question answering to control for spurious correlations (Gat et al., 2020). We shall see in §4.2 that while Q-type performs at chance level, the BERT and LSTM baselines are able to pick up on some spurious correlations. We also evaluate two standard visual question answering architectures, one based on a combination of convolutional and recurrent neural networks (CNN+LSTM), and one attention-based architecture (Hudson and Manning, 2018). The latter performs best on the counting and number comparison tasks in the CLEVR dataset (Johnson et al., 2017) compared with other approaches, such as Bottom-Up-Attention and Top-Down (UpDn) (Anderson et al., 2018), Question-Conditioned Graph (QCG) (Norcliffe-Brown et al., 2018), Bilinear Attention Network (BAN) (Kim et al., 2018), Relation Network (RN) (Santoro et al., 2017) and Recurrent Aggregation of Multimodal Embeddings Network (RAMEN) (Shrestha et al., 2019). We describe each system in detail:

- **Q-type (Antol et al., 2015):** Similar to the "per Q-type prior" method in (Antol et al., 2015), this baseline predicts the most popular answer for each question type.

- **LSTM (Hochreiter and Schmidhuber, 1997):** Question words are embedded as 300-dimensional vector sand fed into an LSTM network. The last hidden state representation is passed into a multi-layer perceptron (MLP) to predict the final answer. All experiments use

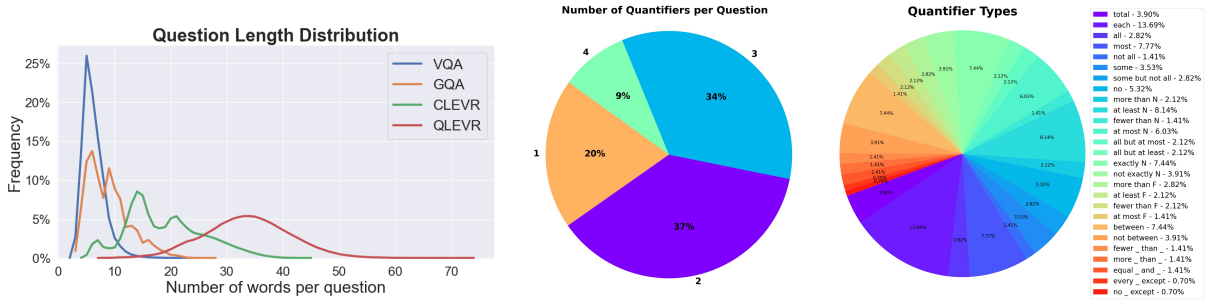


Figure 4: Statistics for our dataset. **Left:** Question length distribution for different popular VQA datasets; most of the QLEVR questions have 30 to 40 words, which is longer than other datasets. **Middle:** Distribution of the number of quantifiers in QLEVR questions. **Right:** Frequency distribution of quantifiers in QLEVR, where *N* stands for *Number* and *F* stands for *Fraction*; *each*, *total*, *no*, *at most N*, *at least N*, *exactly N*, *between*, *not between*, and *not exactly N* are also used in the plane templates, so they appear more frequently. For the quantifiers in the texts of the same question family, we consider each quantifier in square of opposition $\{Q, Q\neg, \neg Q, Q^d\}$ as quantifier Q.

a bi-directional LSTM with 512 units in the hidden layer per direction.

- **BERT (Devlin et al., 2019):** We fine-tune BERT (Devlin et al., 2019) augmented with a sentence-level classification head: The special classification token `[CLS]` is passed to a feed-forward layer and used for sentence class prediction.

- **CNN+LSTM:** The images are encoded using a convolutional neural network and questions as the last hidden state produced by an LSTM network. The convolutional network uses spatial features produced by ResNet-101 (He et al., 2015) pre-trained on ImageNet (Deng et al., 2009). We resize all images to 448x448, and use the final average pooling layer to extracts features of the shape (1, 14, 14, 2048). The question and image features are concatenated and passed to a multi-layered perceptron to predict the final answer.

- **MAC (Hudson and Manning, 2018):** The MAC network is a recurrent attention network, which uses a Memory, Attention, and Composition (MAC) cell in each attention-based reasoning step to learn to perform iterative reasoning processes. MAC learns compositional reasoning directly from the questions and the images in an end-to-end approach. The word vectors have a dimension of 300 and are initialized randomly using a standard uniform distribution. The images are resized to 448x448, and 2048-dimensional features are produced by ResNet-101. The model uses a hidden state size of 512 and a length of 12 MAC cells.

4.2 Analysis by Quantifier Type

Table 3 shows the results of the five methods described in §4.1 on the test set of QLEVR. We make the following observations.

1. Q-type exhibits performance levels around 50% for every quantifier type, showing that the answer distribution of QLEVR is uniform.

2. Text-only LSTM and BERT achieve an average accuracies of 64.6% and 65.8%, respectively. These results suggest that even if the answers of each question family are distributed uniformly, there may still be spurious correlations: Objects with more detailed attribute descriptions may be more likely to get a *false* answer. For example, the question *"Are there more than 3 small blue cubes on the black planes?"* is more likely to get a *false* answer than *"Are there more than 3 blue objects on the black planes?"*).

3. The CNN+LSTM architecture performs better than LSTM on 24 out of 27 quantifier types and on par with BERT; MAC performs better than LSTM on 26 out of 27 quantifier types, better than BERT on 24 out of 27 quantifier types and better than CNN+LSTM on 23 out of 27 quantifier types. In general, however, CNN+LSTM and MAC do not improve much over text-only LSTM and BERT, suggesting that the image features extracted by ResNet-101 contain little information relevant to counting in complex scenes.

4. Accuracies for quantifiers that present thresholds (e.g., *more than N, at least F, etc.*) are

	Q-type	LSTM	BERT	CNN+LSTM	MAC
each	50.0	63.9	65.4	65.3	66.2
total	50.0	63.4	64.7	65.1	66.3
all	50.0	59.5	60.5	60.7	61.3
most	50.0	61.7	63.6	63.5	64.0
not all	50.0	60.7	62.2	61.3	62.6
no	50.0	63.9	64.6	64.6	65.3
some	50.0	58.2	58.9	58.7	58.7
some but not all	50.0	59.9	61.5	61.0	61.7
exactly N	50.0	62.6	63.8	63.5	64.0
not exactly N	50.0	64.9	65.7	65.9	66.7
between	50.0	65.2	66.6	67.0	67.9
not between	50.0	64.4	65.8	66.0	66.0
all but at most	50.0	66.9	68.8	68.2	68.9
all but at least	50.1	62.7	63.7	63.3	65.2

	Q-type	LSTM	BERT	CNN+LSTM	MAC
more than N	50.0	65.5	67.3	67.1	67.2
at least N	50.0	64.7	66.2	65.8	66.5
fewer than N	50.0	66.7	67.6	67.7	68.3
at most N	50.0	64.3	65.4	64.9	65.7
more than F	50.0	69.0	69.6	71.4	71.4
at least F	50.0	67.9	70.1	71.0	72.0
fewer than F	50.0	67.6	68.8	70.2	71.6
at most F	50.0	65.8	68.9	70.3	70.8
every _ except _	50.1	70.9	72.0	70.9	72.1
no _ except _	50.1	78.4	78.1	77.9	78.2
more _ than _	50.0	68.0	69.4	69.1	68.7
fewer _ than _	50.0	68.9	69.9	68.8	69.0
equal _ and _	50.0	61.0	62.2	62.0	62.2
Overall	**50.0**	**64.6**	**65.8**	**65.9**	**66.5**

Table 3: Test set results of baselines and state-of-the-art models on the QLEVR dataset. Models are evaluated for both overall accuracy and accuracy per quantifier type. In quantifier type, *N* stands for *Number* and *F* stands for *Fraction*. Refer to Figure 4 for the number distribution of quantifiers.

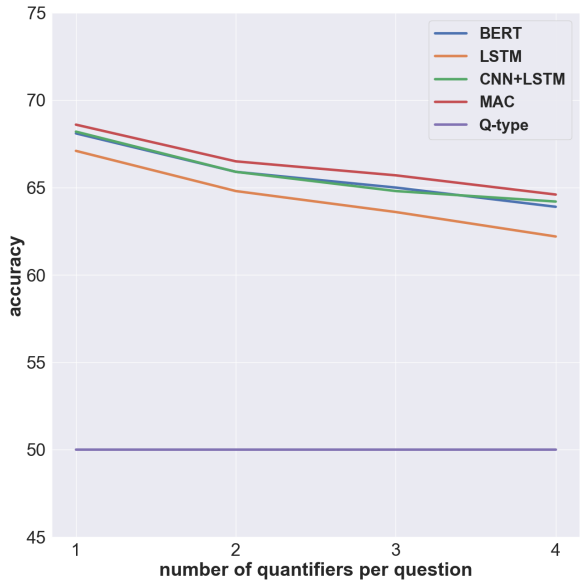


Figure 5: The effect of different number of quantifiers in questions on the accuracy of the answers. Figure 4 shows the distribution of the number of quantifiers in each QLEVR question.

higher than for quantifiers that require a number of objects to match exact values (e.g., *exactly N*).

5. Quantifiers without numerals (e.g., *all, most, not all, some* and *some but not all*) lead to lower accuracies than other quantifiers, showing that reasoning with these quantifiers is harder. This highlights the need for including such quantifiers in challenge datasets to push advancements in visual question answering.

4.3 Analysis

Number of Quantifiers in Questions Figure 5 shows how accuracy varies as the number of quantifiers in the questions increases. The more quantifiers in a question, the more complex its semantics will be.

Number of Planes We also test the visual reasoning abilities of these models by examining error across the number of planes involved in answering the question. Appendix A introduces all the plane templates in our question families. We use the plane template *"on the <PC> <PM> <PS> plane<ps>"* for our analysis, because this template has no influence of quantifiers or spatial relationships in targeting planes. QLEVR test set has 13,612 questions with this plane template. The left graph in Figure 6 shows how the accuracy varies with the increase in the number of target planes that need to be reasoned with. Among the 13,612 questions, 10,288 of them involve a single plane and 3,324 of them involve multiple planes. We can see that for language-only models Q-type, BERT and LSTM, the number of target planes does not significantly affect the accuracy. However, for CNN+LSTM and MAC, questions involving just a single plane are harder to answer than those involving multiple planes. This is because for visual models, planes enable disambiguation and thereby reduce the required reasoning. The right graph in Figure 6 compares accuracy on questions that do not refer to specific planes (*no attribute*), to questions that refer to specific planes (*with attributes*). Among the 13,612 questions, 1,340 questions do not refer to specific planes, whereas 12,272 do. This distinction has little impact on the perfor-

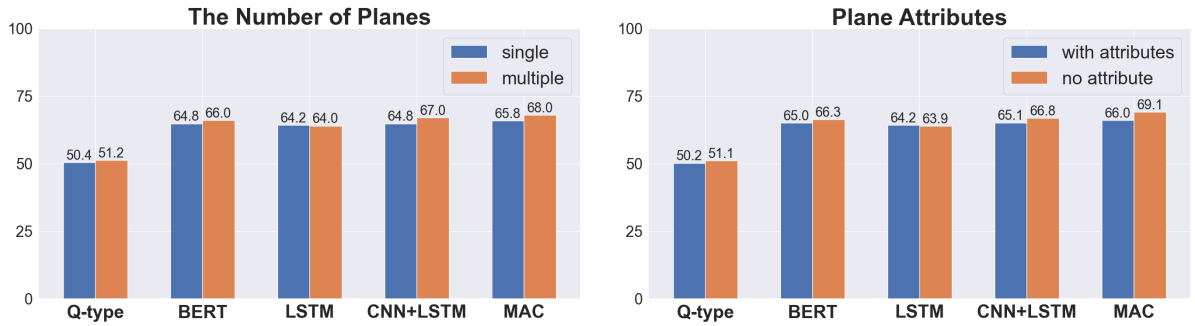


Figure 6: The results of questions contains *"on the <PC> <PM> <PS> plane<ps>"*. **Left**: The effect of different number of target planes on the accuracy of the answers; *single* means that the reasoning process basically only needs to consider one plane in the image, while *multiple* means that multiple planes need to be considered. **Right**: The effect of whether the plane has attribute description on the accuracy of the answer; *no attribute ("on the planes")* means that the reasoning process does not need to consider planes in the image and see the image as a whole, while *with attributes ("e.g., on the wooden plane")* means that specific plane(s) needs to be considered.

mance of our text-only models. For CNN+LSTM and MAC, however, examples in the *no attribute* class exhibit higher accuracies than those in *with attributes*. This, again, shows performance is better when less visual reasoning is required.

5 Discussion

In this paper, we proposed a dataset, which we call QLEVR – for Quantificational Language and Elementary Visual Reasoning. QLEVR probes the ability of visual question-answering systems to reason with quantificational language, including 27 different quantifiers and combinations thereof. It requires complex visual reasoning to locate the specific planes and understand various relationships between objects. We increase the semantic diversity of the questions by negating quantifiers and by using different templates for semantically equivalent questions. Our analysis highlights how challenging such examples are to visual question-answering systems, and we hope that QLEVR will help guide push research horizons in visual question-answering by zooming in on the challenges posed by quantificational language.

One fundamental limitation is that QLEVR only considers English questions, and we plan to extent it to other, typologically unrelated languages. Besides, QLEVR can easily be extended by adding new question families, and questions whose answers are not limited to true or false, e.g., with numbers or attributes as answer types. In addition to the three-dimensional images, we also provide two-dimensional images and scene graphs recording the ground-truth information (see Figure 2). It is also possible to generate questions about 2D images by simply modifying our question families. We hope these two datasets can be used for transfer learning for visual question answering in the future.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: which investigated applications? *Pattern Recognit. Lett.*, 151:325–331.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6):e12417.

Edmund M. Clarke, E. Allen Emerson, and Joseph Sifakis. 2009. Model checking: Algorithmic verification and debugging. *Commun. ACM*, 52(11):74–84.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Lei Fang and Jian-Guang Lou. 2021. Part & whole extraction: Towards a deep understanding of quantitative facts for percentages in text. *ArXiv*, abs/2110.13505.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Advances in Neural Information Processing Systems*, volume 33, pages 3197–3208. Curran Associates, Inc.

Shayan Hassantabar. 2018. Visual question answering : Datasets , methods , challenges and oppurtunities.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Jaakko Hintikka. 1977. Quantifiers in natural languages: Some logical problems ii. *Linguistics and Philosophy*, 1(2):153–172.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NeurIPS*, pages 1682–1690.

Will Norcliffe-Brown, Efstathios Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*.

Iuliia Parfenova, Desmond Elliott, Raquel Fernández, and Sandro Pezzelle. 2021. Probing cross-modal representations in multi-step relational reasoning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 152–162, Online. Association for Computational Linguistics.

Sandro Pezzelle and Raquel Fernández. 2019. Is the red square big? MALeViC: Modeling adjectives leveraging visual contexts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2865–2876, Hong Kong, China. Association for Computational Linguistics.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Shailaja Keyur Sampat, Akshay Kumar, Yezhou Yang, and Chitta Baral. 2021. CLEVR_HYP: A challenge dataset and baselines for visual question answering with hypothetical actions over images. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 3692–3709, Online. Association for Computational Linguistics.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. In *NIPS*.

Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2017. High-order attention models for visual question answering. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer them all! toward universal visual question answering models. In *CVPR*.

Dag Westerståhl. 2012. Classical vs. modern squares of opposition, and beyond.

Supplementary Material

A Question Templates

As described in Section 3, QLEVR question templates are composed of 11 plane templates and 61 object templates randomly paired. In this section we detail the difference between these templates.

Plane Templates. The role of the plane templates is to raise our question for specific planes (regions) in the image through some restrictions (attributes, spatial relations and explicitly restricted quantifier phrases). Basically, the plane templates can generate questions with following types:

- On the white non-geometric planes.
- On the geometric plane with a different shape (color/material) from other planes.
- On the black planes to the left rear of the circular plane.
- On the planes where there are at least 3 red cubes on each plane.
- On the quadrilateral plane where there are at most 5 blue balls.
- On the brown planes where there are between 1 and 4 triangular prisms on each plane.
- On the triangular plane where there is exactly 1 leathery object.
- On the plane where there are not 2 to 4 triangular prisms.
- On the gray planes where there are not exactly 3 items on each plane.
- On the marble plane where there are not any wooden cones.
- On the wooden plane where there is a total of 7 small rubber objects.

To avoid pragmatically odd questions, we ensure that the number of planes obtained by the plane templates with restrictions of spatial relations and explicitly restricted quantifier phrases (e.g. *On the brown planes behind the gray plane*, or *On the brown plane where there are exactly 3 balls*) is less than the number of planes obtained by the templates without these restrictions (e.g. *On the brown planes*) for the same scene graph.

Object Templates. We can use the *operators* representation of the questions templates to analyze model performance on the following forms of reasoning:

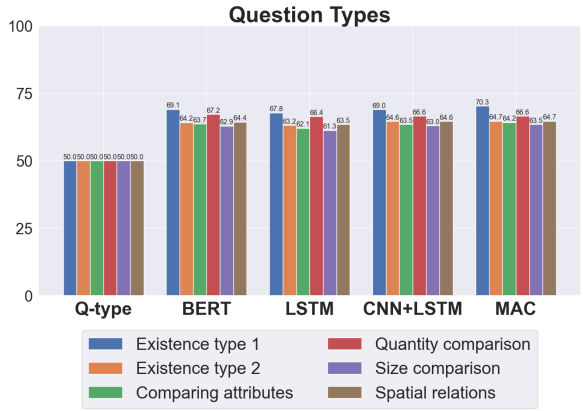


Figure 7: Accuracy per question type on the QLEVR dataset.

- **Existence type 1:** Questions ask whether a certain type of quantifier-restricted object exists on one or some specific planes (*e.g., "Whether all the cyan cubes* [Plane Template]*?"*).
- **Existence type 2:** Questions ask whether a certain type of quantifier-restricted object exists in a certain direction of a unique object (*e.g., "*[Plane Template]*, are there fewer than 3 balls behind the cyan cube?"*).
- **Comparing attributes:** Questions ask whether two types of quantifier-restricted objects have the same value for some attributes (*e.g., "*[Plane Template]*, is there any small cylinders that has the same color as most leathery tetrahedrons?"*).
- **Quantity comparison:** Questions compare the size of two sets of objects (*e.g., "*[Plane Template]*, are there more big blocks than rubber balls?"*).
- **Size comparison:** Questions ask which of two quantifier-restricted objects has a larger size (*e.g., "*[Plane Template]*, some red cones are larger than some but not all of the metal cones; is it right?"*).
- **Spatial relations:** Questions involves the spatial relationship between objects (*e.g., "*[Plane Template]*, are there more big blocks in front of the yellow cylinder than rubber balls to the left rear of the small block?"*).

Figure 7 shows the performance on above question types. As can be seen, MAC outperforms

other models on most question types. The only exception is: on quantity comparison task, BERT performs slightly better than MAC, showing that MAC has better reasoning ability in complex scenes. Questions of *Existence type 1* obtain better results than *Existence type 2* for vision-language model CNN+LSTM and MAC, suggesting that the position relationship between object and plane is easier to be inferred by the models than the spatial relationship between the objects. For questions of *Quantity comparison*, MAC and CNN+LSTM performs on par with LSTM, suggesting that the image features extracted by ResNet-101 may contain little information related to counting in complex scenes.

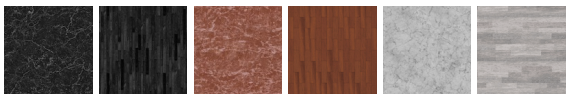
B 3D Modeling and Design

Figure 8 shows the materials and object models made through Blender (Community, 2018), as well as the performance of different colors on these materials. Two different materials of leather, marble, and wood were made respectively to further enrich the diversity of objects in the dataset. The images of the plane materials were made by modifying the images under CC0 1.0 Universal.[1] Note that after the overall scene rendering, objects of certain materials will produce different effects according to the color and material of the plane in contact with them, as well as the position of the camera and lights.

C Example Images and Questions

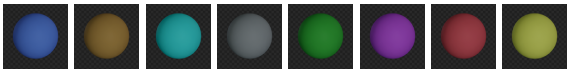
The remaining pages show some images and questions generated by the combination of our different plane templates and object templates. Each question is annotated with its answer and contained quantifiers, where *N* stands for *Number*, *F* stands for *Fraction* and *O* stands for *Object*.

(a) object shapes

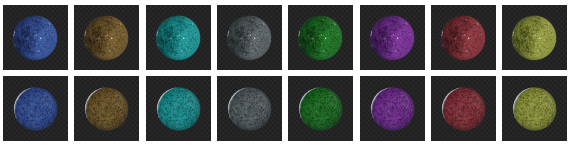


(b) materials and colors of the planes

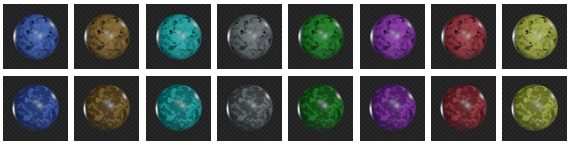
(c) metal



(d) rubber



(e) leather

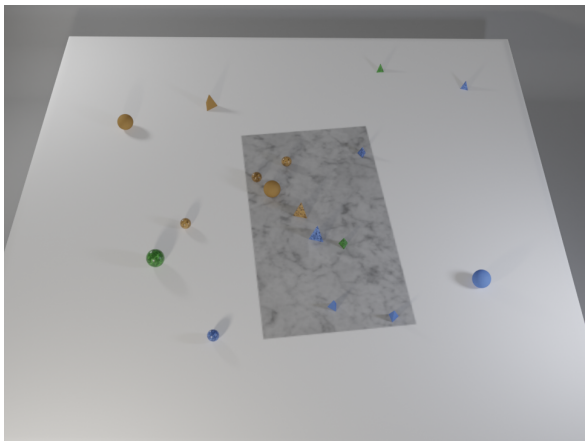


(f) marble

(g) wood

Figure 8: From left to right, the object shapes in (**a**) are cone, cube, cylinder, pentahedron, sphere, triangular prism, and tetrahedron; the plane attributes in (**b**) are black marble, black wood, brown marble, brown wood, gray marble and gray wood; the colors in (**c**) ~(**g**) are blue, brown, cyan, gray, green, purple, red and yellow.

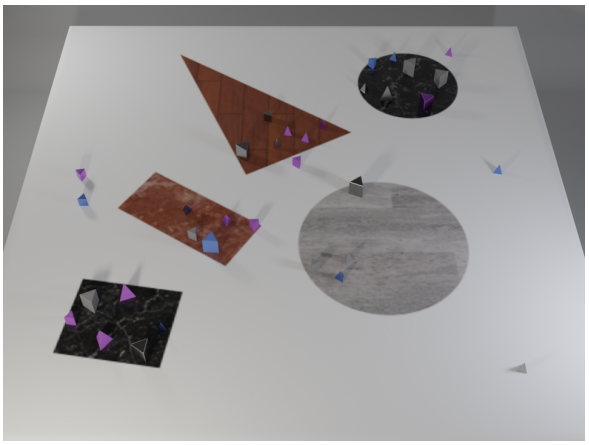
[1] Creative Commons - CC0 1.0 Universal



Question: Whether all the large brown objects are on the white plane?
Answer: False
Quantifiers: all

Question: Some large rubber tetrahedron is not on the gray marble plane; is it right?
Answer: True
Quantifiers: not all (some ¬)

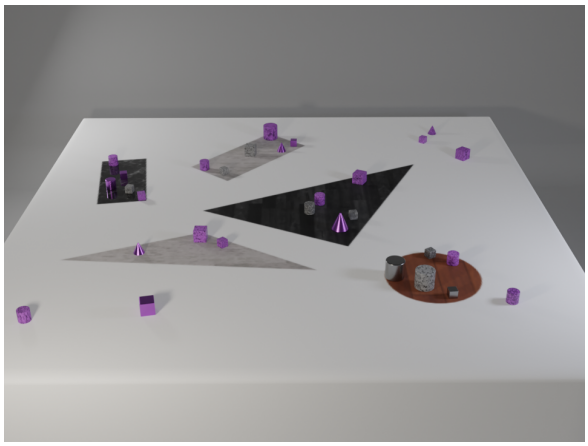
Question: It is not the case that all the big blue rubbery spheres are not on the gray rectangular plane; is it right?
Answer: False
Quantifiers: some (¬ all ¬)



Question: It's not the case that some large purple metallic triangular prism is on the planes where there are 9 items in total; is it right?
Answer: True
Quantifiers: total, no (¬ some)

Question: Whether some but not all of the large purple rubber objects are on the marble planes where there are 4 blue objects in total?
Answer: False
Quantifiers: total, some but not all

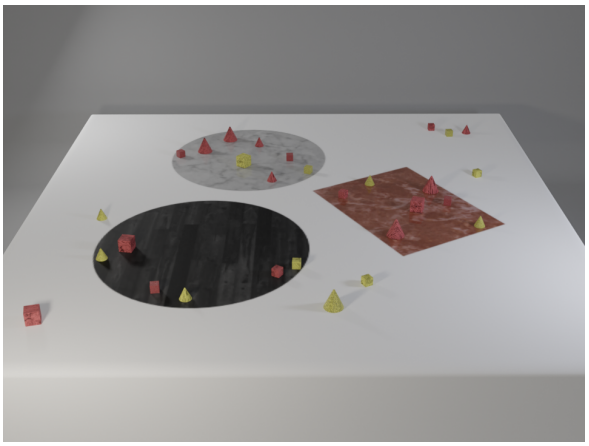
Question: Are there at most 3 small blue objects on the dappled planes where there are 5 big triangular prisms in total?
Answer: True
Quantifiers: total, at most N



Question: All the big wooden blocks but at least 2 are not on the planes where there are exactly 2 cylinders on each plane; is it right?
Answer: True
Quantifiers: each, exactly N, at least N (all but at least N ¬)

Question: It is not the case that at most 2 wood cylinders are on the quadrilateral plane where there are exactly 2 purple wood cylinders; is it right?
Answer: False
Quantifiers: each, exactly N, more than N (¬ at most N)

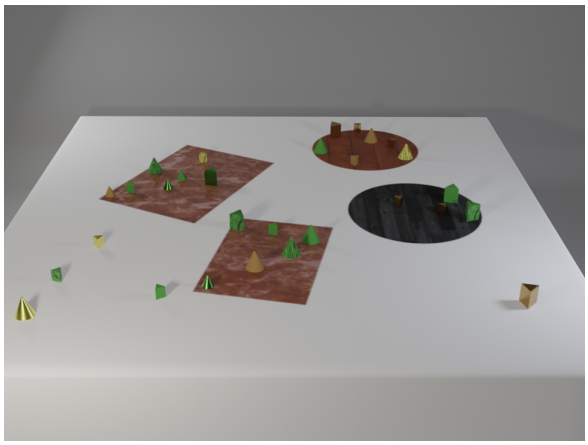
Question: Are there fewer than 2 small purple dappled cylinders on the planes where there is exactly 1 purple block on each plane?
Answer: True
Quantifiers: each, exactly N, fewer than N



Question: All the tiny red wood objects but 1 are not on the non-white plane where the shape of the plane is different from that of other planes; is it right?
Answer: True
Quantifiers: exactly N (all but N ¬)

Question: Are there between 1 and 3 small red leathery cubes on the circular plane to the left rear of the brown quadrilateral plane?
Answer: True
Quantifiers: between

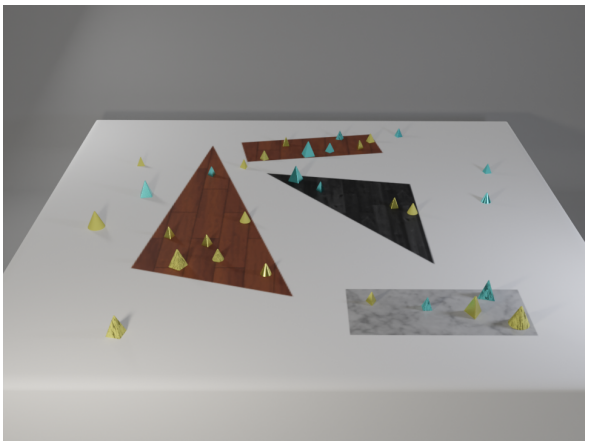
Question: All the red leather objects but at most 3 are on the geometric plane where the material of the plane is different from that of other planes; is it right?
Answer: False
Quantifiers: all but at most N



Question: More than two thirds of the green objects are on the brown marble plane to the left front of the black wood plane; is it right?
Answer: False
Quantifiers: more than F

Question: Are most brown metallic objects on the geometric plane on the right side of the brown wooden round plane?
Answer: False
Quantifiers: most

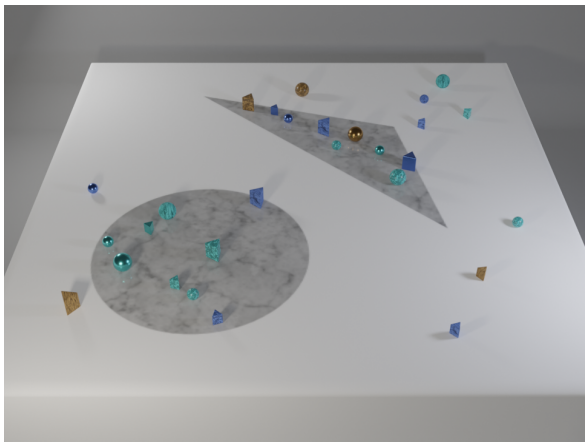
Question: It is not the case that fewer than 4 brown metallic triangular prisms are not on the non-white plane where the color of the plane is different from that of other planes; is it right?
Answer: True
Quantifiers: all but at least N (¬ fewer than N ¬)



Question: Fewer than three-quarters of the small yellow objects are on the wood three-sided plane where there are not any large metallic square-based pyramids; is it right?
Answer: True
Quantifiers: no (¬ any), fewer than F

Question: It is not the case that fewer than 11/15 of the metallic items are on the planes where there are 0 tiny rubbery tetrahedrons on each plane; is it right?
Answer: True
Quantifiers: each, no (0), at least F (¬ fewer than F)

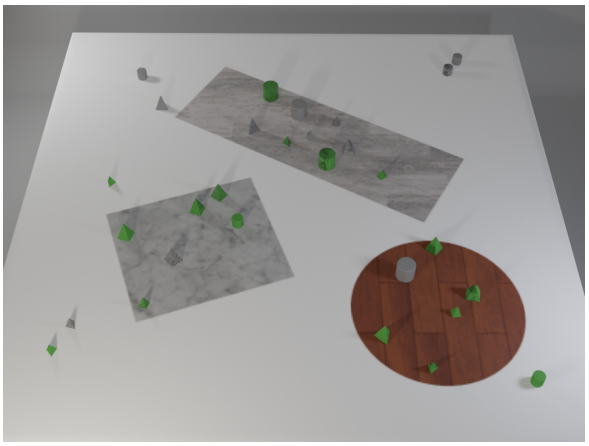
Question: At most 7/8 of the yellow rubber triangular pyramids are on the planes where there is no big yellow wooden cone on each plane; is it right?
Answer: False
Quantifiers: each, no, at most F



Question: On the gray marble plane where there are between 3 and 5 tiny cyan objects, is there any wooden triangular prism that has the same size as most metallic objects?
Answer: False
Quantifiers: between, some (any), most

Question: On the gray plane where there are between 1 and 4 cyan triangular prisms, 2 to 5 cyan items are the same material as most small items; is it right?
Answer: True
Quantifiers: between, between, most

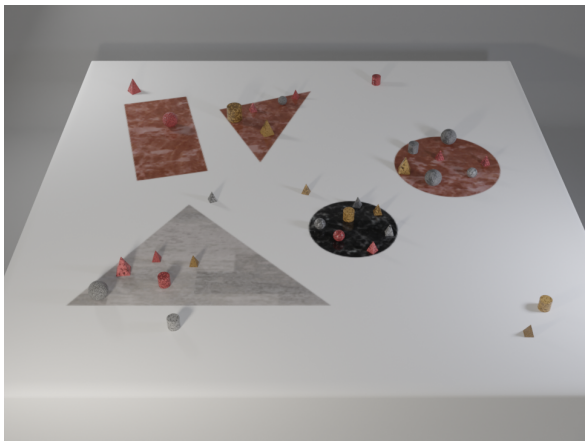
Question: On the planes where there are between 3 and 6 cyan items on each plane, are there exactly 3 small triangular prisms that have the same color as most small metallic spheres?
Answer: True
Quantifiers: each, between, exactly N, most



Question: On the gray quadrilateral plane where there are not between 1 and 3 small gray items, are all the large items but at most 1 the same color as most leather items?
Answer: True
Quantifiers: not between, all but at most N, most

Question: On the plane where there are not between 2 and 4 green leather objects, fewer than a half of the gray cylinders are the same size as most gray rubbery objects; is it right?
Answer: False
Quantifiers: not between, fewer than F, most

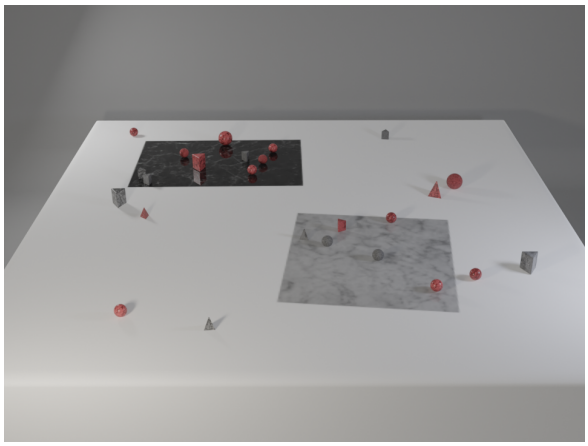
Question: On the planes where there are not between 0 and 3 big objects on each plane, more than five twelfths of the big gray objects have the same shape as most gray rubber objects; is it right?
Answer: False
Quantifiers: each, not between, more than F, most



Question: On the planes where there are not exactly 2 spheres on each plane, at most 4 red leather objects are the same shape as most objects; is it right?
Answer: True
Quantifiers: each, not exactly N, at most N, most

Question: On the triangular plane where there is not exactly 1 small marbled square-based pyramid, at least 2 red marbled items have the same size as most red items; is it right?
Answer: False
Quantifiers: not exactly N, at least N, most

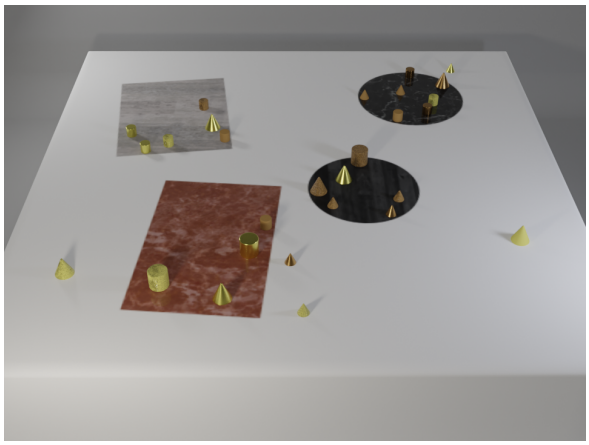
Question: On the marbled planes where there is at least 1 large marbled item on each plane, at least 2 large items are not the same material as most gray items; is it right?
Answer: True
Quantifiers: each, at least N, all but at least N (at least N ¬), most



Question: On the dappled planes, is there the same number of tiny objects on the left side of the big sphere and tiny red spheres right of the red leathery triangular prism?
Answer: False
Quantifiers: equal O_1 and O_2

Question: On the white non-geometric plane, are there fewer big objects to the right rear of the big red dappled object than tiny spheres in front of the big red dappled object?
Answer: True
Quantifiers: fewer O_1 than O_2

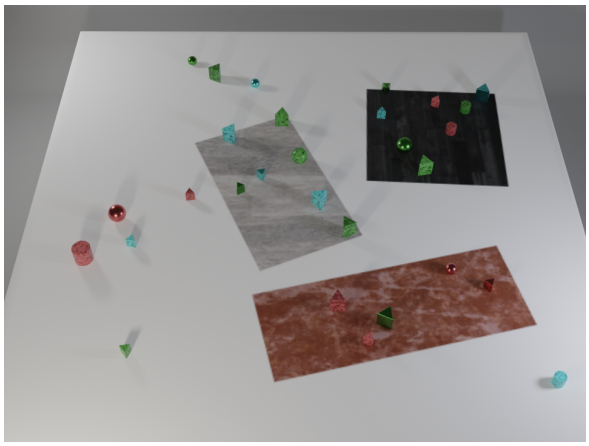
Question: On the planes, is the number of triangular prisms to the right front of the red leathery tetrahedron greater than the number of leathery tetrahedrons on the left side of the large marble sphere?
Answer: True
Quantifiers: more O_1 than O_2



Question: On the planes where there are more than or equal to 2 cylinders on each plane, are there more brown metallic cones than brown leathery cylinders?
Answer: False
Quantifiers: each, at least N, more O_1 than O_2

Question: On the wood plane where there is not exactly 1 small yellow leathery object, is the number of small objects less than the number of brown leathery cones?
Answer: False
Quantifiers: not exactly N, fewer O_1 than O_2

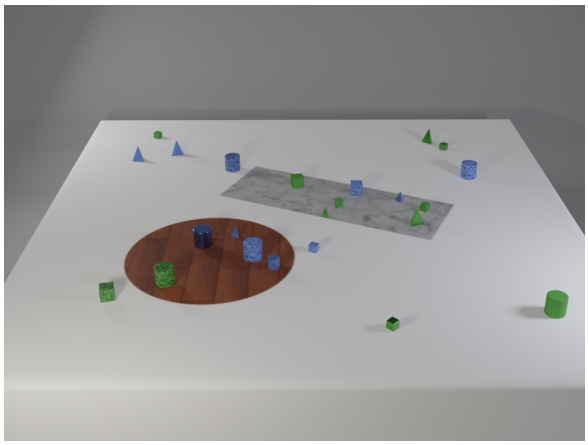
Question: On the planes where there are no fewer than 2 big cones on each plane, is the number of circular cylinders the same as the number of big circular cylinders?
Answer: True
Quantifiers: each, at least N, equal O_1 and O_2



Question: On the planes where there are no more than 3 tiny marbled objects on each plane, all the tiny marbled objects are in front of the red metallic triangular prism; is it right?
Answer: True
Quantifiers: each, at most N, all

Question: On the plane where there are at most 4 marble items, is there any tiny ball in front of the tiny red metal three-sided prism?
Answer: False
Quantifiers: at most N, some

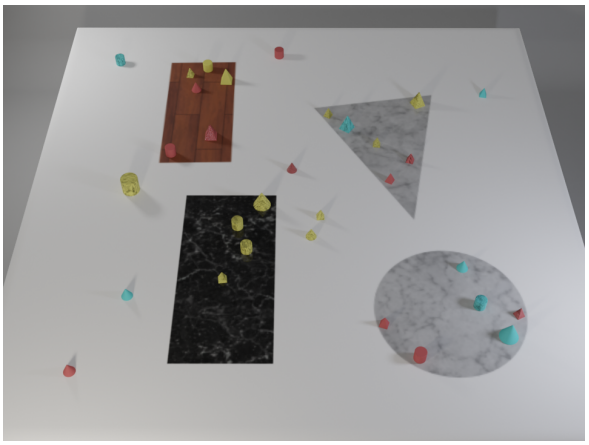
Question: On the wood plane where there are fewer than or equal to 3 small marble objects, it is not the case that no large marble triangular prism is not on the left side of the green sphere; is it right?
Answer: True
Quantifiers: at most N, not all (¬ no ¬)



Question: On the planes where there is exactly 1 small blue marbled object on each plane, every green metal object is not behind the green rubber cylinder; is it right?
Answer: False
Quantifiers: each, exactly N, no (every ¬)

Question: On the plane where there is not exactly 1 big green metal tetrahedron, some but not all of the marble objects are to the right of the blue tetrahedron; is it right?
Answer: True
Quantifiers: not exactly N, some but not all

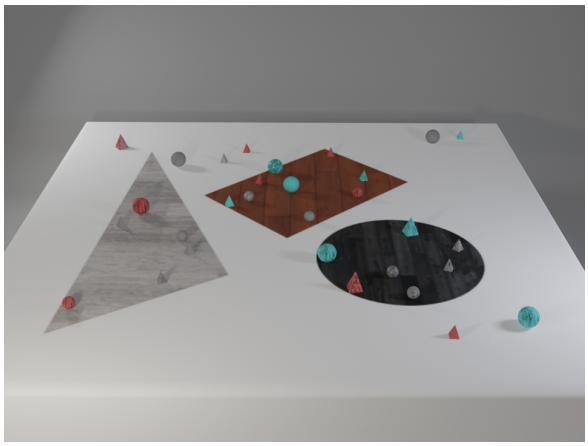
Question: On the planes where there is a total of 5 big green items, all the tiny green blocks but at most 1 are not right rear of the big green marble block; is it right?
Answer: False
Quantifiers: total, at most N (all but at most N ¬)



Question: On the planes where there are at least 2 cyan items on each plane, all the tiny cones but at least 4 are not right rear of the tiny red rubbery pentahedron; is it right?
Answer: False
Quantifiers: each, at least N, at least N (all but at least N ¬)

Question: On the quadrilateral plane where there is at most 1 yellow circular cylinder, it is not the case that at most 1 red rubber object is left rear of the leather object; is it right?
Answer: False
Quantifiers: at most N, more than N (¬ at most N)

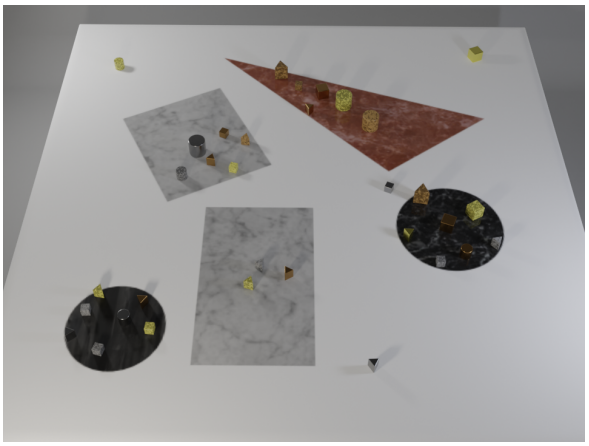
Question: On the geometric plane whose material is different from that of other planes, it is not the case that all the red items but at least 2 are not right front of the tiny yellow cylinder; is it right?
Answer: True
Quantifiers: fewer than N (¬ all but at least N ¬)



Question: On the planes where there are not any small red rubber objects on each plane, there are exactly 2 small gray pentahedrons left rear of the large cyan pentahedron; is it right?
Answer: False
Quantifiers: each, no (¬ any), exactly N

Question: On the planes where there is not exactly 1 wood ball on each plane, are there 1 to 3 wood objects to the left front of the gray marbled square-based pyramid?
Answer: False
Quantifiers: each, not exactly N, between

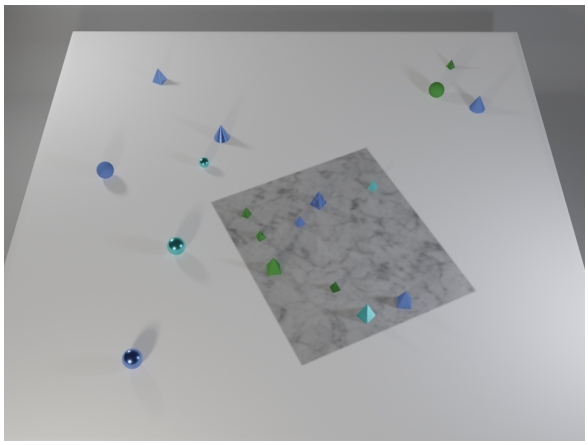
Question: On the wooden planes where there are not between 1 and 3 red rubber items on each plane, at most 3 small items are not on the right front side of the big red sphere; is it right?
Answer: True
Quantifiers: each, not between, all but at most N (at most N ¬)



Question: On the planes where there are between 1 and 3 tiny gray items on each plane, are all the tiny items but at least 5 right of the large gray metal circular cylinder?
Answer: False
Quantifiers: each, between, all but at least N

Question: On the round plane to the left of the brown dappled three-cornered plane, most items are behind the small yellow dappled block; is it right?
Answer: True
Quantifiers: most

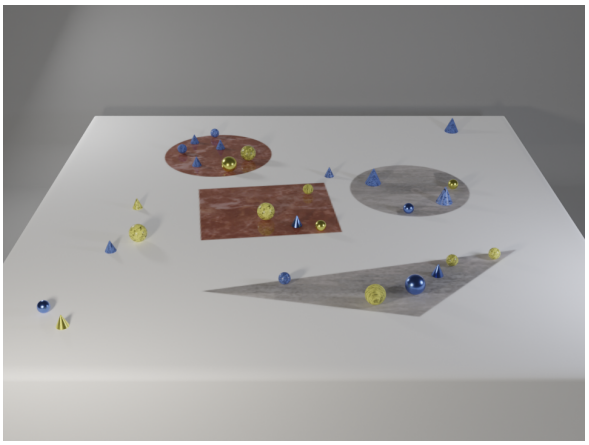
Question: On the marble planes where there is a total of 7 tiny triangular prisms, it is not the case that fewer than 2/3 of the big items are to the left of the tiny gray cube; is it right?
Answer: False
Quantifiers: total, at least F (¬ fewer than F)



Question: On the gray marble plane, it is not the case that more than 5/8 of the pentahedrons are to the right of the tiny metallic pentahedron; is it right?
Answer: True
Quantifiers: at most F (¬ more than F)

Question: On the rectangular plane, all the blue wood square pyramids are larger than some rubber square pyramid; is it right?
Answer: True
Quantifiers: all, some

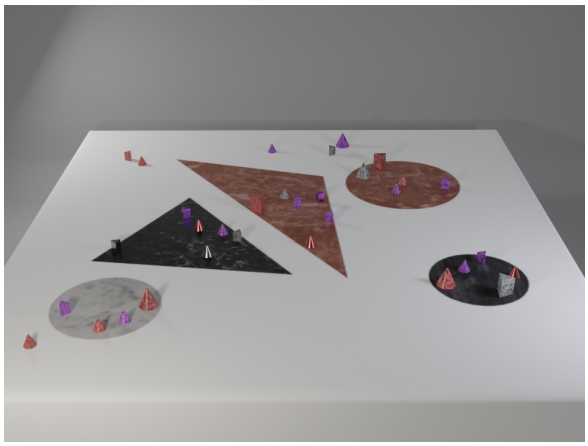
Question: On the non-geometric plane, are all the green square-based pyramids smaller than some but not all of the square-based pyramids?
Answer: True
Quantifiers: all, some but not all



Question: On the planes where there are 0 large metal balls on each plane, it is not the case that no blue cone is larger than some but not all of the yellow cones; is it right?
Answer: False
Quantifiers: each, no (0), some (¬ no), some but not all

Question: On the planes where there are not exactly 3 tiny blue objects on each plane, exactly 3 blue cones are larger than at least 2 blue wood cones; is it right?
Answer: False
Quantifiers: each, not exactly N, exactly N, at least N

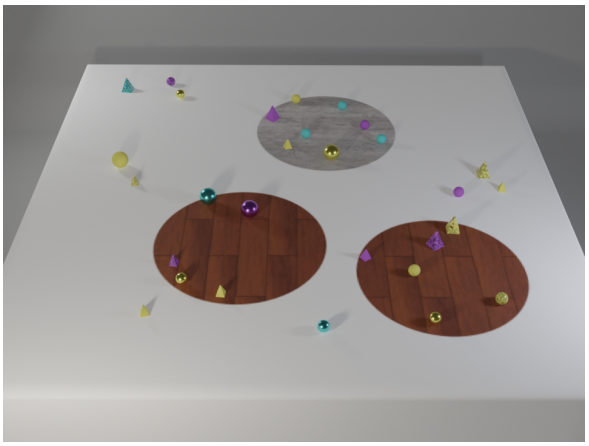
Question: On the planes where there are not between 1 and 4 tiny marbled spheres on each plane, all the spheres but at least 2 are larger than at most 1 yellow sphere; is it right?
Answer: True
Quantifiers: each, not between, at least N (all but at least N ¬), more than N (¬ at most)



Question: On the circular planes where there are exactly 2 big objects on each plane, 1 to 3 marbled cones are smaller than more than 3/7 of the red marbled cones; is it right?
Answer: True
Quantifiers: each, exactly N, between, more than F

Question: On the brown plane where there are between 1 and 3 red marbled objects, at least one-ninth of the marbled cones are larger than at least 2 cones; is it right?
Answer: True
Quantifiers: between, at least N, at least F

Question: On the marble planes where there is a total of 5 marble items, all the cones but at least 2 are smaller than fewer than 3/4 of the gray cones; is it right?
Answer: False
Quantifiers: total, at least N (all but at least N ¬), at least F (¬ fewer than F)



Question: On the geometric plane whose color is different from that of other planes, all yellow spheres but 1 are smaller than fewer than 1 or more than 3 spheres; is it right?
Answer: True
Quantifiers: exactly N (all but N ¬), between (¬ between)

Question: On the wooden planes where there are no fewer than 2 tiny balls on each plane, every item except the marble ball is not a tiny item; is it right?
Answer: False
Quantifiers: each, at least N, no _ except (every _ except ¬)

Question: On the brown plane where there is no more than 1 dappled square pyramid, no objects except the metal ones are not square pyramids; is it right?
Answer: True
Quantifiers: at most N, every _ except (no _ except ¬)