A Templates from CLEVR dataset generator

Here are the relevant templates from the CLEVR dataset generator for each function word. Words in brackets are options. The placeholders stand for: -S- is for shapes or other referent nouns; -M- are sets of adjectival modifiers (between 0 and 3 for size, color, material); -P- is for a property noun ("color", "size", "material", "shape"); -R- is for a spacial relation (e.g. left of, right of, behind, above). Note that the function words *behind* and *in front of* are considered relations in the CLEVR generative model and thus don't appear in the templates but can replace any -R- placeholder. They can therefore appear in some of these templates as well as many others.

And

Are there an equal number of -M- -S-s and -M2- -S2-s?
Are there the same number of -M- -S-s and -M2- -S2-s?
Are there the same number of -M2- -S2-s [that are] -R- the -M- -S- and -M3- -S3-s?
Are there an equal number of -M2- -S2-s [that are] -R- the -M- -S- and -M3- -S3-s?
Are there an equal number of -M2- -S2-s [that are] -R- the -M- -S- and -M4- -S4-s [that are] -R2- the -M3- -S3-?
Are there the same number of -M2- -S2-s [that are] -R- the -M- -S- and -M4- -S4-s [that are] -R2- the -M3- -S3-?
Do the -M- -S- and the -M2- -S2- have the same -P-?
Are the -M- -S- and the -M2- -S2- made of the same material?
Do the -M2- -S2- [that is] -R- the -M- -S- and the -M3- -S3- have the same -P-?
Do the -M- -S- and the -M3- -S3- [that is] -R- the -M2- -S2- have the same -P-?
Do the -M2- -S2- [that is] -R- the -M- -S- and the -M4- -S4- [that is] -R2- the -M3- -S3- have the same -P-?
Are the -M2- -S2- [that is] -R- the -M- -S- and the -M3- -S3- made of the same material?
Are the -M- -S- and the -M3- -S3- [that is] -R- the -M2- -S2- made of the same material?
Are the -M2- -S2- [that is] -R- the -M- -S- and the -M4- -S4- [that is] -R2- the -M3- -S3- made of the same material?
How many -M3- -S3-s are [both] -R2- the -M2- -S2- and -R- the -M- -S-?
What number of -M3- -S3-s are [both] -R2- the -M2- -S2- and -R- the -M- -S-?
What is the -P- of the -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-?
What -P- is the -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-?
How big is the -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-?
There is a -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-; what is its -P-?
There is a -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-; what -P- is it?
There is a -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-; how big is it?
The -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S- is what color?
What is the -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S- made of?
The -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S- is made of what material?
There is a -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S-; what [material] is it made of?
The -M3- -S3- that is [both] -R2- the -M2- -S2- and -R- the -M- -S- has what shape?

Or

How many [things/objects] are [either] -M- -S-s or -M2- -S2-s?
What number of [things/objects] are [either] -M- -S-s or -M2- -S2-s?
How many -P- [things/objects] are [either] -M- -S-s or -M2- -S2-s?
What number of -P- [things/objects] are [either] -M- -S-s or -M2- -S2-s?
How many [things/objects] are [either] -M2- -S2-s [that are] -R- the -M- -S- or -M4- -S4-s [that are] -R2- the -M3- -S3-?
What number of [things/objects] are [either] -M2- -S2-s [that are] -R- the -M- -S- or -M4- -S4-s [that are] -R2- the -M3- -S3-?
How many -S3-s are [either] -M- -S-s or -M2- -S2-s?
What number of -S3-s are [either] -M- -S-s or -M2- -S2-s?
How many [things/objects] are [either] -M2- -S2-s [that are] -R- the -M- -S- or -M3- -S3-s?
What number of [things/objects] are [either] -M2- -S2-s [that are] -R- the -M- -S- or -M3- -S3-s?

How many [things/objects] are [either] -M- -S-s or -M3- -S3-s [that are] -R- the -M2- -S2-?
What number of [things/objects] are [either] -M- -S-s or -M3- -S3-s [that are] -R- the -M2- -S2-?

More
Are there more -M2- -S2-s [that are] -R- the -M- -S- than -M4- -S4-s [that are] -R2- the -M3- -S3-?
Are there more -Z- -C- -M- -S-s than -Z2- -C2- -M2- -S2-s?
Are there more -M2- -S2-s [that are] -R- the -M- -S- than -M3- -S3-s?

Fewer
Are there fewer -M2- -S2-s [that are] -R- the -M- -S- than -M4- -S4-s [that are] -R2- the -M3- -S3-?
Are there fewer -M- -S-s than -M2- -S2-s?
Are there fewer -M2- -S2-s [that are] -R- the -M- -S- than -M3- -S3-s?

B MAC model hyperparameters

Here are the hyperparameters of the models we used for the experiments reported in the main paper. All models were trained on an Nvidia RTX 3080 GPU, taking about 15-20 hours including the probe evaluations every epoch. Like in the original paper (Hudson & Manning, 2018), models are trained using an Adam optimizer (Kingma & Ba, 2015). Word vectors were initialized randomly using a standard uniform distribution based on a random seed. The models used a variational dropout of 0.15 across the network. We also used ELU as our non-linearity type.
number of epochs: 25
learning rate: 0.0001, if performance on the validation set didn't improve from an epoch to the next, then the learning rate got reduced using a decay rate of 0.5.
batch size: 64
number of MAC cell layers: 4
hidden size: 512 throughout network
embedding size: 300
image features output size: 1024
random seed: [0:4]

C On models learning SAME

In initial versions of these experiments, we also included probes for the relational adjective *same*. Though the word *same* is part of the CLEVR vocabulary, the word *different* is not. So instead of having a pair of opposing words for this reasoning skill, we considered how models learnt to interpret *same* in contexts where objects did or did not share attributes. We tried two probe versions, one which tested the comparison of object sets and a second simpler one which required one-to-one object comparison. We chose not to include these results in the main paper as these probes proved to be either too hard or too easy for the models to answer, leading to overall uninformative results as to the learning dynamics of these models.

C.1 The set comparison SAME probe results

This version of the SAME probe required models to consider sets of objects of varied sizes rather than to compare two specific referents. Questions were of the form 'Are the *X*s the same *property*', X being all possible referring expressions for objects within the CLEVR universe and *property* being either *color*, *shape*, *size* or *material*. We matched possible properties to referents such that the value of these properties was not explicitly given by the referent terms (e.g. If the referents were *brown sphere* and *metal cube*, then the only possible value for property was *size*, since the referent terms already mention the color, material, or shape of one of the objects). This first template presupposed that there were at least two *X*s. After finding all images for each question where this presuppositions was satisfied and then sampling 10 images if there were more than 10 available, this probe had a total of 1,170 image-question pairs.

In order for the answer to a question to be 'yes' in the case of the first template all Xs in the image had to have the same value for the *property* listed in the question, otherwise the answer was 'no'. We used the 'scene' metadata file associated with each image to verify whether this condition was met.

We ran this version with random seeds [0-2] before deciding to change to a different SAME question template, so the results presented here are based on 3 random runs rather than 5 like the rest of the experiments in the paper. We plot the mean F1 scores and standard deviation across runs.

Figure S1 shows the overall performance of the models on this version of the SAME probe. The models' performance is quite low.

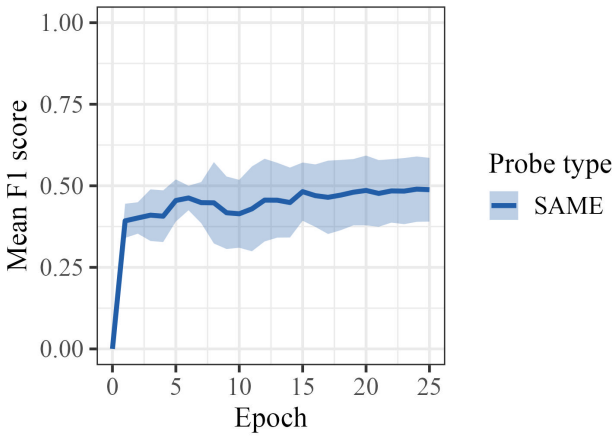

Figure S1: Mean F1 score on previous SAME probes overall. Shading represents standard deviation across 3 models.

We can also look at results by answer type in Figure S2. Again, it is clear that models really struggle to answer these questions as in either case they are only really considering two possible answers: "yes" or "no".

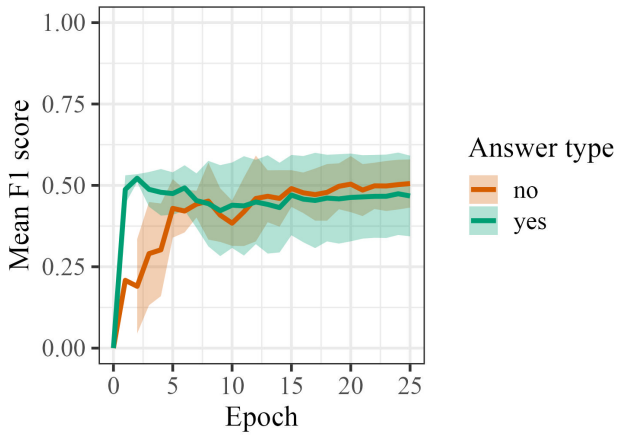

Figure S2: Mean F1 score on previous SAME probes by answer type.

This initial version of the SAME probe may have been too different of a use of the word *same* from those the model was exposed to during training, which could explain the models poor performance.

C.2 The one-to-one comparison SAME probe results

SAME probes Given that models struggled with probe questions of the form 'Are the *X*s the **same** *property*?', we opted to try 'Are the *X* and the *Y* the **same** *property*?', where *property* is one of the following: *shape*, *color*, *material*, *size*. This second template presupposes that there are exactly one *X* and one *Y*. After finding all images for each question where this presuppositions was satisfied and then sampling 10 images if there were more than 10 available, the second SAME probe had 38,580 question-image pairs.

In order for the answer to a question to be 'yes', *X* and *Y* had to share the same value for the given *property* for a 'yes' answer, otherwise the answer was 'no'.

We ran this version with random seeds [0-4], so the results presented here are based on 5 random runs like the rest of the experiments in the paper. We plot the mean F1 scores and standard deviation across runs.

Figure S3 shows the overall mean F1 scores of models on this SAME probe using the newer two referent comparison template instead. The models have almost perfect scores.

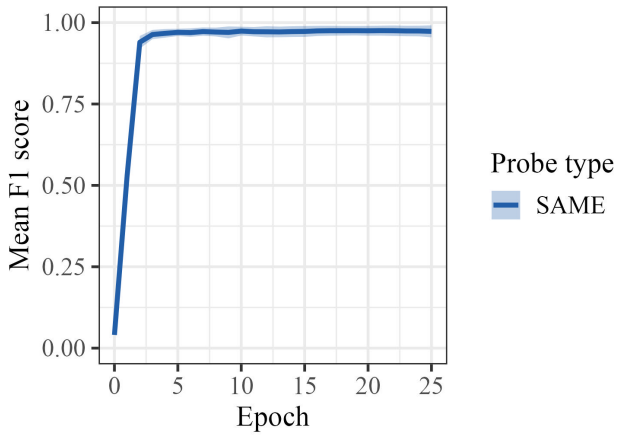

Figure S3: Mean F1 score on SAME probe overall. Shading represents standard deviation across 5 models.

When we considered models' performance as a function of the answer type expected, again we found that they score almost perfectly in either context, Figure S4.

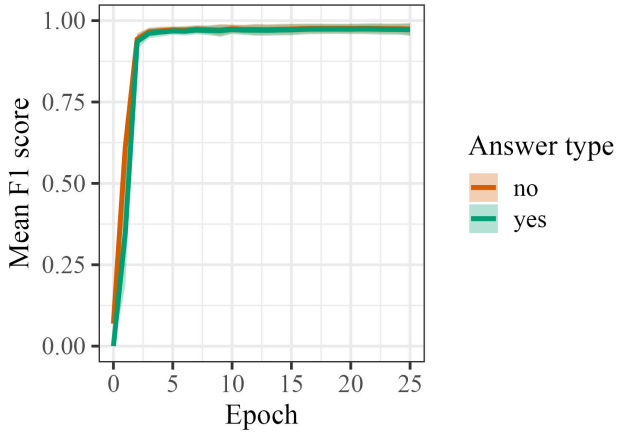

Figure S4: F1 score on SAME probe by answer type. Shading represents standard deviation across 5 models.

Thus, though the models may struggle to generalize using *same* to reason about items in a set, they learn to quickly use it to compare two unique referents. Additionally, they show no difference between correctly answering questions in contexts where these two referents share the *same property* versus contexts where they do not.

References

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations.*

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Procceedings of the 3rd International Conference for Learning Representations.*