# Supplementary: Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules

In this supplementary document, we discuss the following:

1. Structure of Query Generator (section 1)
2. Query-focused soft masking of capsules (section 2)
3. Interpreting attention visualizations (section 3)
4. Qualitative results (section 4)
5. Further results and analysis (section 5)
    - Results for best vs. last reasoning step
    - Results comparison w.r.t. question type
    - Results comparison w.r.t. reasoning type
    - Reduction in parameters
    - Impact of opacity parameter $\alpha$ on grounding

## 1. Structure of Query Generator

The query generator is a recurrent module proposed by MAC [2], also used in SNMN [1], to obtain question-related reasoning operation $q_t$ at each step t. The recurrent module essentially takes its output from previous timestep $t-1$ i.e., $q_{t-1}$ along with question features $f_s$ and $f_w$ as sentence and word embeddings respectively, and generate attention over question words at current timestep $t$. More specifically, the query generator applies a time-step dependent linear transformation on $f_s$ and combines it with the previous reasoning operation $q_{t-1}$ as follows:

$$u = W_2([W_1^t(f_s) + b_1; q_{t-1}]) + b_2 \quad (1)$$

where, $W_1^t$ and $W_2$ are $d \times d$ and $d \times 2d$ transformation matrices respectively. To generate attention over question words $f_w$, following is done:

$$a_w = softmax(W_3(u \odot f_w) + b_3) \quad (2)$$

where, $W_3$ is a $2d \times d$ matrix and $a_w$ are the attention scores over question words. Finally, $q_t$ is obtained by taking the weighted sum over $l$ question words using $a_w$.

$$q_t = \sum_{w=1}^{l} a_w \cdot f_w \quad (3)$$

**Note:** We refer the readers to the works MAC[2] and SNMN[1] for additional details about these systems.

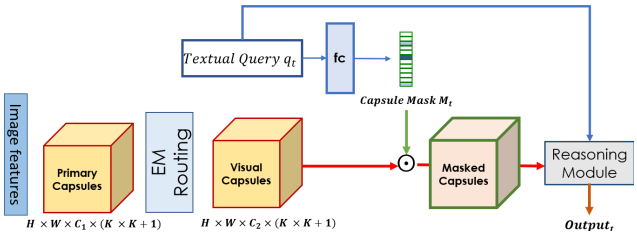


Figure 1. Query-focused Soft Masking of capsules: first, we transform convolutional image features into visual capsules. Then, we use a question-based feature called textual-query $q_t, t = 1, 2, ..., T$ to select relevant capsules by soft masking capsules which are irrelevant to the textual query. These masked capsules are then input to a reasoning module for further processing. Reasoning module depends on how a VQA system is designed to perform reasoning for answering the question. We show integration of this module into two VQA systems: MAC [2] and SNMN [1] as shown in the main paper, figure 2 and section 4.

## 2. Query-focused soft masking of capsules

The proposed query-focused soft masking of capsules is a generic module which can be integrated seamlessly into existing VQA systems. First, the convolutional image features are transformed into the visual capsules as explained in the main paper. Then, a question-based feature called textual-query $q_t$ is used to select relevant capsules by soft masking capsules which are irrelevant to the textual query. For a multihop VQA system with T hops, $q_t$ is the textual-query at timestep $t$, where $t \in \{1, 2, ..., T\}$. These masked capsules are then input to a reasoning module for further processing. What a reasoning module is, depends on how a VQA system is designed to perform reasoning for answering the question. We show integration of this capsule module into two VQA systems: MAC [2] and SNMN [1] as shown in the main paper, figure 2 and section 4. Adding the proposed capsule module results in significant improvement in grounding in both baseline methods.

## 3. Interpreting attention visualization

We start with a short explanation of the qualitative visualization of attention steps. To gain insight into how the

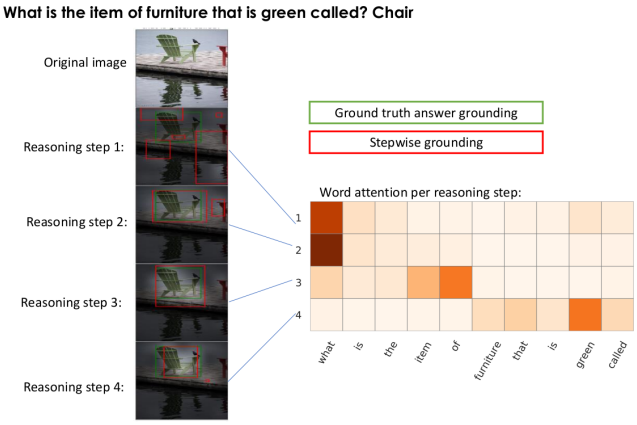


Figure 2. To visualize the attention maps produced at each reasoning step, we display the original image and question pair, followed by the four reasoning steps, where each attended area is highlighted and marked by a red bounding box. To compare the attended area (grounding) to the ground truth, we also display the ground truth bounding box in green. The attention for each step is guided by the respective word attentions at each reasoning step.

VQA system works in combination with grounding, we visualize both, the grounding within the image as well as the respective word attentions within the question. As shown in figure 2, we display the original image and the question, followed by the output of the different reasoning steps, *e.g.*, four in case of MAC, with attended regions highlighted and marked by a red bounding box. The ground truth bounding box is shown in green.

Note that in the paper, we report grounding results for the best overlap between grounding and ground truth, independent of the reasoning step. A non-overlapping bounding box at another reasoning step, *e.g.*, at step one, does not influence the final result. We choose this best-of-all metric, because, especially for SNMN and SNMN-Caps, it is possible that the model might look at the correct grounding at an intermediate step and not necessarily at the last step. See "Best vs. Last" in section 5 for more details.

Additionally, we also show the respective word attentions for each reasoning step. As shown in figure 2, each row shows the attention potential of each reasoning step. Note that for efficient usage of space, this matrix is flipped in all other visualizations and the reasoning steps are shown column wise. In the case shown here, one can see that the first two reasoning steps are based on attentions on the word "What", the third reasoning step is based on "item of", and the last reasoning step attends to "green". Note that the term "furniture" is not used to answer the question.

## 4. Qualitative Results

**GQA** Figure 3 shows more qualitative comparison between MAC and MAC-Caps. Both MAC and MAC-Caps were trained with T=4 number of reasoning steps. For instance in figure 3's first example (columns 1-2), MAC attends to question words such as "is" at T=1. At T=2, attention is spread, including the word "green" but also paying similar amount of attention to words "is" and "the". At T=3 and 4, MAC gives most of its attention to words "is" and "what" respectively. When looking into the question-to-image attention visualization, we observe MAC attending to corners in the image, also in the second step, thus the impact of the attention on green is rather marginal, except for the final reasoning step (last row) where it has an overlap with the correct object based on the attention to "What". MAC-Caps, on the other hand, attends to the word "what" at T=1,2 with the attention on different regions on image (including "green chair"). At T=3, MAC-Caps is looking at words "item of", and finally attends to the second half of the question at T=4 (with most attention to the word "green"). MAC-Caps performs much better in localization of "green chair" in the image (column 2, rows 4 and 5).

**CLEVR-Answers** For CLEVR-Answers, we show step-by-step attention visualization for SNMN and SNMN-Caps. Both models have been trained with T=9 reasoning steps, i.e., each question may use up to 9 reasoning steps if that many steps are required to answer the question. For visualization, we follow [1] and remove $No\_Op$ if a question takes fewer than 9 steps to produce an answer. See figure 4 for qualitative analysis. We keep the grounding results from the reasoning step giving best grounding output in terms of F1-score. We observe SNMN-Caps is consistent in producing better attention maps covering most of the ground truth boxes. Besides, SNMN-Caps attends to correct objects (or to nothing) based on the input textual query. For instance, in figure 4, column 4, row 5: for the textual query "tiny gray shiny", SNMN-Caps produces an empty grounding map, whereas, SNMN attends to a wrong object (cyan cube, column 3, row 4) for the same textual query. Similarly, in the last example (columns 5-6), for textual query "right", SNMN misses one object on the right; SNMN-Caps, however, attends to all objects on the right with better overlap to the ground truth grounding boxes. Figure 5 shows examples where no grounding evidence was available for the question resulting in an empty map. SNMN fails to produce attention maps with uniform attention in such cases resulting in false positive detections. SNMN-Caps, however, has learned that it can attend to nothing, thus, correctly generating empty maps in the final step.

## 5. Further results and analysis

**Best vs. Last** We first compare the baselines and their respective capsule-based variants for best-of-all metric com-

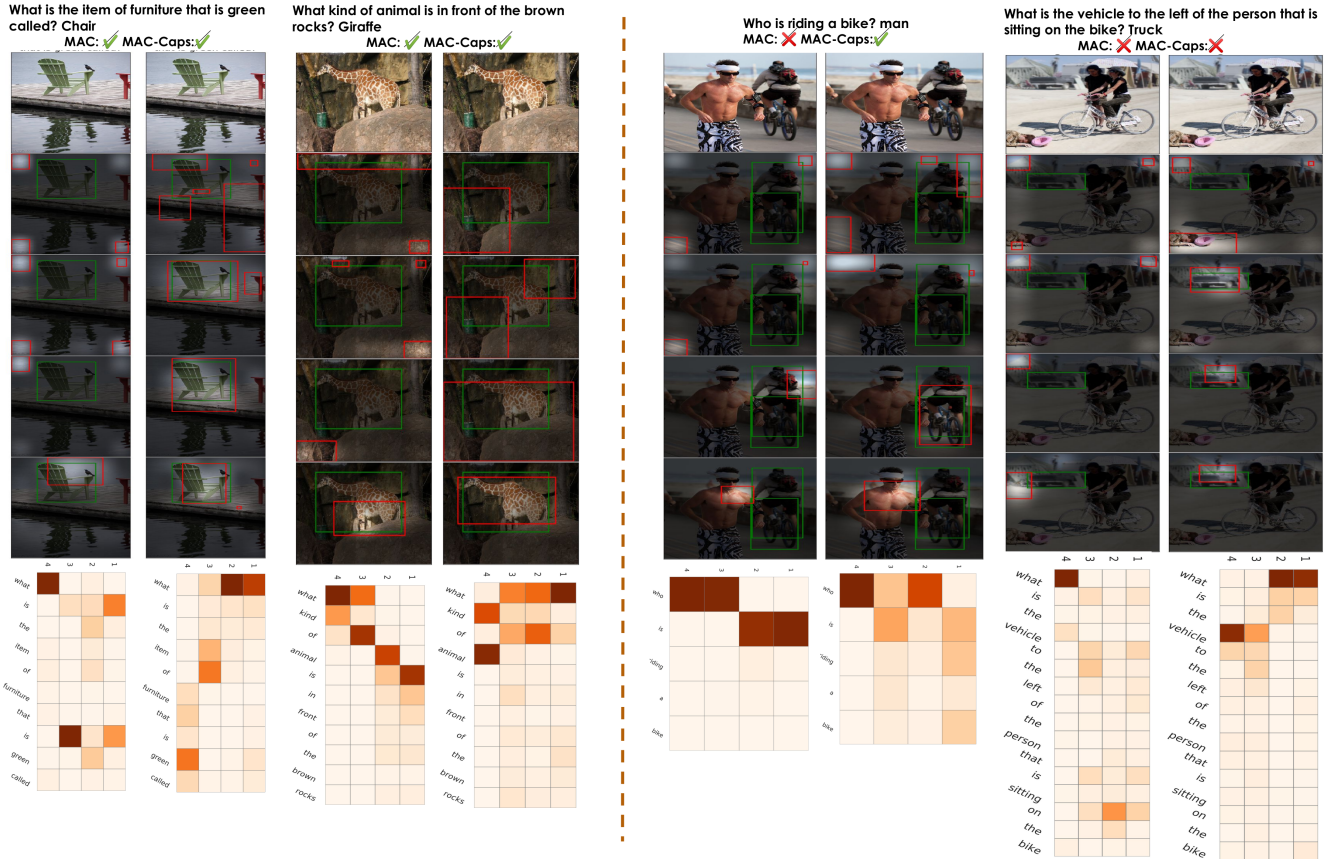


Figure 3. Attention visualizations for MAC on GQA dataset for success (columns 1-4) and failure cases (columns 5-8). Column 1 shows results for MAC, column 2 shows results for MAC-Caps and the same order is followed onwards. Row 1 shows input image, rows 2-5 are attention visualizations for each reasoning step (T=4) with ground truth (green boxes) and detected grounding objects (red boxes), followed by attention on question words for each reasoning step. MAC-Caps attends to the correct boxes with better overlap; for instance, see example 1, MAC tends to attend corner regions of the image, whereas MAC-Caps starts looking on different image regions and is able to quickly locate "green" chair. Similar trend is observed among all examples shown here. The attention on question words is also improved for MAC-Caps with more attention to relevant words when compared to MAC. Refer to section 4 for further details and discussion. Best viewed in color.

| **CLEVR-Answers** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Overlap | | | | | | IOU | | | | | |
| | | | P | | R | | F1 | | P | | R | | F1 | |
| **Method** | **T** | Acc. | best | last | best | last | best | last | best | last | best | last | best | last |
| MAC [2] | 4 | **97.70** | 24.92 | 11.64 | 56.27 | 29.63 | 34.55 | 16.72 | 13.99 | 6.45 | 33.50 | 16.40 | 19.73 | 9.25 |
| MAC-Caps | | 96.79 | **47.04** | **31.22** | **73.06** | **57.03** | **57.23** | **40.35** | **23.97** | **12.24** | **39.06** | **22.15** | **29.71** | **15.77** |
| MAC [2] | 6 | 98.00 | 30.10 | 12.98 | 52.14 | 25.77 | 38.24 | 17.26 | 12.59 | 5.06 | 23.62 | 10.03 | 16.42 | 6.73 |
| MAC-Caps | | **98.02** | **48.49** | **28.12** | **79.75** | **54.74** | **60.31** | **37.15** | **29.03** | **13.23** | **47.63** | **25.47** | **36.07** | **17.41** |
| MAC [2] | 12 | **98.54** | 28.66 | 9.47 | 53.27 | 21.33 | 37.27 | 13.11 | 8.50 | 2.62 | 18.11 | 5.89 | 11.57 | 3.62 |
| MAC-Caps | | 97.88 | **50.90** | **26.22** | **94.61** | **60.26** | **66.19** | **36.54** | **27.72** | **9.27** | **49.84** | **20.97** | **35.62** | **12.86** |
| | **#param** | | | | | | | | | | | | | |
| SNMN [1] | 7.32M | 96.18 | 52.87 | 44.13 | 67.03 | 56.64 | 59.12 | 49.61 | 37.81 | 29.51 | 47.50 | 37.32 | 42.11 | 32.96 |
| SNMN-Caps | 6.94M | **96.66** | **73.81** | **63.25** | **78.13** | **67.64** | **75.91** | **65.37** | **50.58** | **40.54** | **51.80** | **41.96** | **51.18** | **41.24** |

Table 1. **Best vs. Last:** Comparison with baseline systems on CLEVR-Answers validation set for grounding from best vs. last reasoning step. Best: the reasoning step in which the respective models achieving the best F1-score, last: final reasoning step is used to evaluate for grounding. MAC-Caps and SNMN-Caps are the variants with the proposed soft masked capsules. For MAC, results are shown with varying reasoning steps. SNMN uses T=9. See section 5 for details. Numbers are reported in percentages.

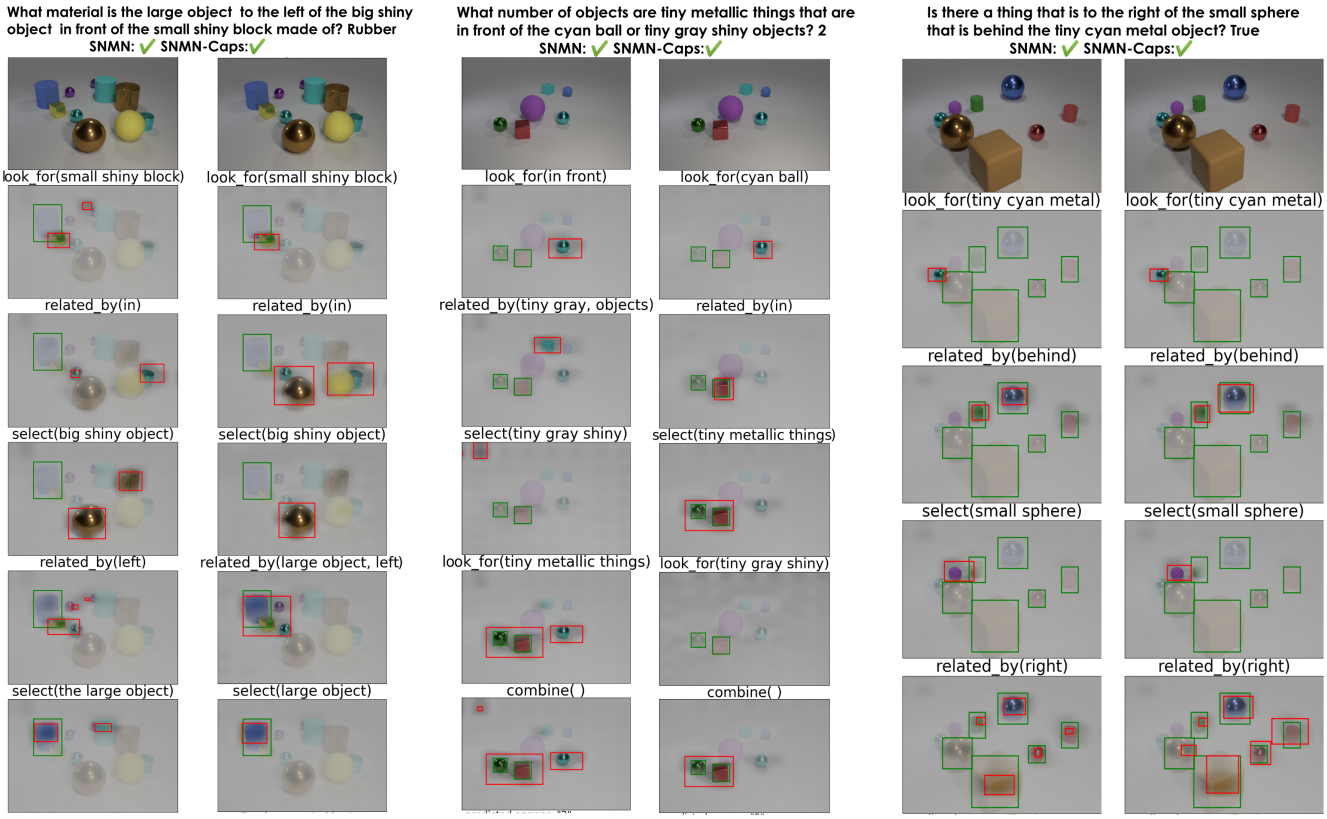


Figure 4. Attention visualizations on the CLEVR-Answers dataset for both SNMN and SNMN-Caps at each reasoning step. We present the question and the input image given to the network in the first row. Each subsequent row is a reasoning step. The reasoning module with the highest weight and the question words with the highest attention (textual query) are displayed above each reasoning step. The red bounding boxes are the detections produced after post-processing the attention maps. The green bounding boxes are the ground-truths, which are displayed for the final reasoning step and the reasoning step in which the respective models achieve the best F1 score. In general, SNMN-Caps produces better groundings, especially at the final reasoning step.

| | | **GQA** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Overlap | | | | | | IOU | | | | | |
| | | P | | R | | F1 | | P | | R | | F1 | |
| **Method** | **Grd. GT** | best | last | best | last | best | last | best | last | best | last | best | last |
| MAC | Q | 19.75 | 10.79 | 30.69 | 16.38 | 24.04 | 13.01 | 2.88 | 1.39 | 4.36 | 2.09 | 3.46 | 1.67 |
| MAC-Caps | Q | **37.77** | **17.39** | **63.65** | **28.10** | **47.41** | **21.49** | **5.39** | **1.87** | **8.65** | **2.96** | **6.64** | **2.29** |
| MAC | FA | 22.43 | 13.62 | 31.35 | 18.63 | 26.15 | 15.74 | 3.30 | 1.80 | 4.48 | 2.42 | 3.80 | 2.06 |
| MAC-Caps | FA | **41.53** | **19.69** | **63.00** | **28.58** | **50.06** | **23.31** | **6.14** | **2.27** | **8.85** | **3.23** | **7.25** | **2.67** |
| MAC | A | 5.61 | 5.05 | 27.36 | 24.44 | 9.31 | 8.37 | 0.92 | 0.76 | 4.46 | 3.70 | 1.52 | 1.27 |
| MAC-Caps | A | **11.95** | **5.46** | **62.56** | **27.90** | **20.07** | **9.13** | **2.32** | **0.97** | **11.91** | **4.94** | **3.88** | **1.62** |
| MAC | All | 25.01 | 15.23 | 30.48 | 18.03 | 27.47 | 16.51 | 3.66 | 1.97 | 4.28 | 2.28 | 3.95 | 2.11 |
| MAC-Caps | All | **46.06** | **22.16** | **62.30** | **27.98** | **52.96** | **24.73** | **7.03** | **2.53** | **8.72** | **3.10** | **7.79** | **2.79** |

Table 2. **Best vs. Last:** Results on GQA validation set for MAC with T=4 for grounding from best vs. last reasoning step. Best: the reasoning step in which the respective models achieving the best F1-score, last: final reasoning step is used to evaluate for grounding. Results are based on grounding of objects referenced in the question (Q), full answer (FA), short answer (A), as well as combined grounding of question and answer (All). We consistently outperform MAC in all metrics. When evaluating for a certain grounding label type, other detected objects are treated as false positives. VQA accuracy is reported in the main paper (table 3). Numbers are reported in percentages.

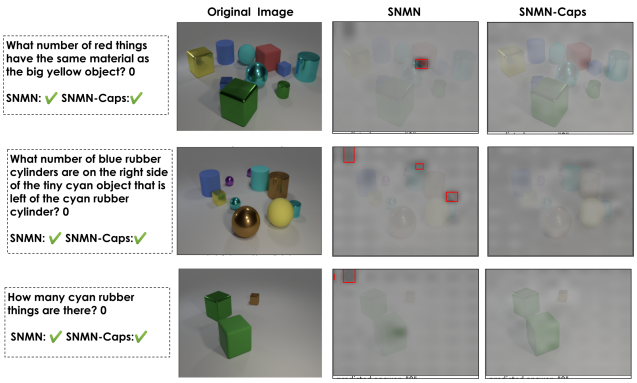


Figure 5. Attention visualizations on the CLEVR-Answers dataset for both SNMN and SNMN-Caps for the final reasoning step on examples with empty grounding maps. Each row displays a data sample with grounding output from SNMN and SNMN-Caps. First two columns show the input image with respective question and answer, third column shows grounding output from SNMN, last column is the grounding from SNMN-Caps. SNMN baseline mostly pays high attentions on regions (red boxes) including objects whereas SNMN-Caps learns to predict nothing by producing uniform attention when the ground truth is supposed to be an empty map.

pared to the performance of the grounding of the last reasoning step (in table 1 and table 2 ). We find that scores degrade significantly for the final reasoning step compared to keeping the best score, because a model may attend to the correct answer regions at any step (including intermediate steps) and process them further to produce an answer. Nonetheless, we see similar gain in grounding score over baselines even when results from the last step are considered. For MAC on **CLEVR-Answers**, we observe the gap is reduced between best results for the baseline vs. last results for SNMN-Caps. For recall, SNMN-Caps (last) is always better than SNMN (best). For MAC=4, we observe that SNMN-Caps (last) is better than the best score for SNMN on all metrics in terms of overlap. These observations indicate that adding capsules has improved grounding both for intermediate reasoning steps as well as the final step. We report results for the best grounding step in the main paper. For **GQA**, we evaluate for grounding of relevant objects in the question (Q), sentence-based full answer (FA), single word answer (A), and for all objects in question-answer pair (All). For all grounding label types, we notice a similar drop in grounding scores if the attention map from the final reasoning step is considered for both MAC and MAC-Caps. We still outperform the MAC baseline model in terms of overlap and IOU; however, the gap between F1-scores for IOU is reduced compared to keeping the best score. This is not surprising because when visualizing the attentions produced by MAC, we notice that it usually looks at the correct answer grounding regions in the last step. MAC-Caps, on the other hand, performs better regardless of which reasoning step (best or last) is used for evaluation.

**Results comparison w.r.t. question type.** Table 3 shows comparison of SNMN and SNMN-Caps models on CLEVR-Answers dataset for different question types, e.g., count, exist, and so on. Although we observe a boost in grounding scores with the proposed capsules module on all question families; we notice that question type $exist$ and $compare\_number$ are the most challenging question types. When looking at the IOU scores, SNMN and SNMN-Caps yield F1-scores of 24.75 vs. 35.83 and 36.40 vs. 41.00, respectively. Both of these question families have boolean (yes/no or true/false) answers, i.e., chance of failure is 50%. For question type $exist$, the lower grounding performance can be attributed to the boolean nature of these questions. For $compare\_number$ question type, the reasoning operation (hence attention) is split among multiple reasoning steps which also leads to a lower grounding score. In terms of overlap, $count$ and $query\_attribute$ seems to be easier questions for grounding where we observe F1-scores of 80.26 and 83.67 respectively. Overall, with our approach, we obtain 17.49% and 9.39% improvement in F1-scores for overlap and IOU respectively.

**Results comparison w.r.t. reasoning type.** Table 4 shows results breakdown of SNMN-Caps on CLEVR-Answers dataset w.r.t. reasoning type–a fine-grained breakdown of grounding results. CLEVR has compositional questions which may need a varying number of reasoning operations to answer them, e.g., a "two hop" question requires two reasoning hops to be answers. We observe the lowest grounding F1-scores obtained on $compare\_integer$ both in terms of overlap and IOU. This is consistent with the previous observation that question type of $compare\_number$ ($compare\_integer$ in reasoning type) is more challenging for grounding relative to the grounding of other reasoning operations.

**Reduction in Parameters.** Since the capsule representation is more compact than the original image features (16 capsules require $d = 16 \times 16 + 16 = 272$ dimensional vector representation, as opposed to the $d = 512$ dimensional feature maps generated from convolutions in the baseline systems), operations within the reasoning modules require fewer parameters. When extending SNMN with 16 capsules, the number of learned parameters reduces by 15.67% (from 7.32M to 6.2M parameters); in MAC with T=4, there is a 7.86% reduction (17.66M to 16.28M). Even with 16 capsules, capsules perform really well in the grounding task (see table 4 in the main paper) indicating that capsules inherently have an advantage over its convolutional variants even with fewer parameters. For grounding, we see similar performance in MAC with 16 capsules using less parameters. However, for MAC, we use C=32 for network length T=4, C=24 for T=6, and C=32 for T=12 because of the best

| Reason type | Method | Overlap | | | IOU | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| Count | SNMN | 47.28 | 75.43 | 58.13 | 29.43 | 45.65 | 35.79 |
| | SNMN-Caps | 73.99 | 87.69 | 80.26 | 42.35 | 45.72 | 43.97 |
| Exist | SNMN | 26.91 | 74.15 | 39.49 | 16.90 | 46.24 | 24.75 |
| | SNMN-Caps | 54.83 | 87.70 | 67.48 | 29.56 | 45.49 | 35.83 |
| Comp. Num. | SNMN | 46.27 | 64.95 | 54.04 | 30.82 | 44.44 | 36.40 |
| | SNMN-Caps | 58.51 | 78.23 | 66.95 | 35.38 | 48.74 | 41.00 |
| Comp. Attr. | SNMN | 84.57 | 47.64 | 60.95 | 73.57 | 40.97 | 52.63 |
| | SNMN-Caps | 92.62 | 55.08 | 69.08 | 82.85 | 45.40 | 58.66 |
| Query Attr. | SNMN | 62.93 | 75.39 | 68.60 | 46.94 | 56.69 | 51.36 |
| | SNMN-Caps | 78.61 | 89.42 | 83.67 | 57.47 | 66.81 | 61.79 |
| Overall | SNMN | 52.10 | 66.48 | 58.42 | 37.38 | 47.38 | 41.79 |
| | SNMN-Caps | 73.81 | 78.13 | 75.91 | 50.58 | 51.80 | 51.18 |

Table 3. Results comparison w.r.t reasoning type on CLEVR validation set (for best reasoning step). Numbers are reported in percentages.

| Reasoning type | Method | Overlap | | | IOU | | |
|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 |
| Zero hop | | 76.39 | 84.03 | 80.03 | 55.23 | 59.53 | 57.30 |
| One hop | | 69.15 | 86.89 | 77.02 | 47.18 | 58.56 | 52.26 |
| Two hop | | 73.10 | 90.21 | 80.76 | 51.25 | 63.95 | 56.90 |
| Three hop | | 74.59 | 92.02 | 82.39 | 53.36 | 67.81 | 59.72 |
| Single OR | | 88.31 | 88.33 | 88.32 | 59.44 | 50.24 | 54.46 |
| Single AND | | 69.80 | 88.11 | 77.89 | 46.97 | 61.03 | 53.09 |
| Same relate | | 66.76 | 88.78 | 76.21 | 39.39 | 51.02 | 44.46 |
| Comparison | | 92.62 | 55.08 | 69.08 | 82.85 | 45.40 | 58.66 |
| Compare integer | | 58.51 | 78.23 | 66.95 | 35.38 | 48.74 | 41.00 |
| Overall | | 73.81 | 78.13 | 75.91 | 50.58 | 51.80 | 51.18 |

Table 4. Results breakdown w.r.t reasoning type on CLEVR dataset for SNMN-Caps (for best reasoning step). Questions with reasoning types *same_relate* and *compare_integer* are more challenging (IOU F1-score is $< 45\%$) for answer grounding than other reasoning types. See section 5, paragraph 3 for more analysis. Numbers are reported in percentages.

scores on VQA task.

**Impact of opacity parameter $\alpha$ on grounding.** To obtain grounding detections from the attention maps, we introduce an opacity parameter $\alpha$. Specifically, [1] used $\alpha$=3 to suppress uniform attention regions by upscaling of opacity in those regions. For SNMN-Caps, we observed some capsules were activated on the background, particularly when no object of interest is found in the image. Although, we find that SNMN-Caps has high recall when compared to SNMN, increasing $\alpha$ improved precision of attention maps which led to the increased F1-score for both overlap and IOU. We perform same post processing on SNMN and SNMN-Caps to report numbers. We noticed the scores for SNMN remain unaffected by parameter $\alpha$ unless increased to a very high value. Figure 5 shows impact of opacity on grounding results.

**Capsules can model background.** While studying the capsules' behavior, we observe that our model has an advantage on samples with no ground truth boxes. More specifically, we take SNMN-Caps model trained with CLEVR-Answers and used train-val split for this study. Interestingly, we find that some capsules are focusing more on the background. When carefully examined for examples where the grounding output should be an empty map, we find that capsules are looking at the background for 677 out of these 1586 samples rather than focusing on any object, and are better than the baseline for 83.17% of such cases. The orig-

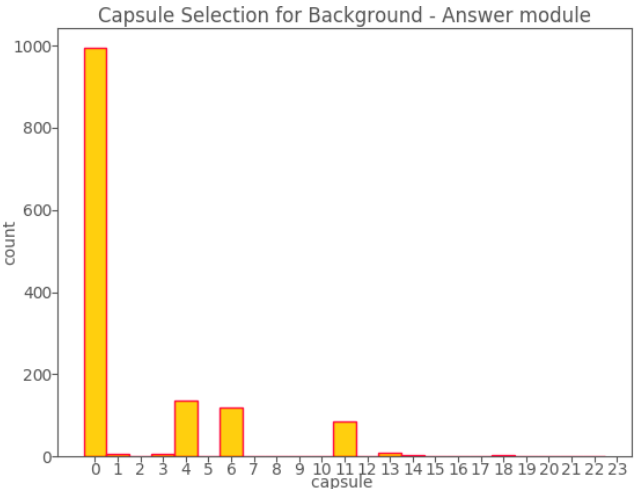


Figure 6. Background capsules selected by the $Answer$ module in SNMN-Caps on CLEVR-Answers train-val set for questions with empty grounding maps. MAC-Caps was trained with 24 capsules. X-axis shows the capsule number, and Y-axis shows the frequency (count of questions) of a particular capsule being selected. For each question in this subset, we select the highest probability capsule in the answer module (probability scores are generated with a soft-masking layer, where, capsules with less probability scores are considered masked or not selected). As we can see that capsule 0 is contributing the most for questions with empty grounding maps. See 5 for further details.

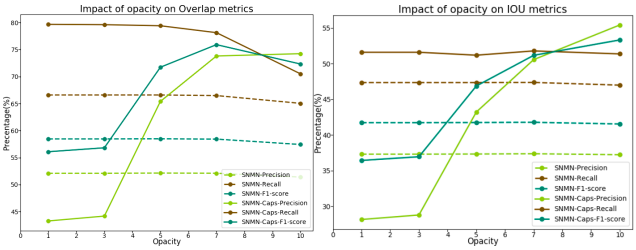


Figure 7. Impact of opacity parameter $\alpha$ on grounding results for SNMN and SNMN-Caps. Left: impact of opacity parameter $\alpha$ on overlap in terms of precision, recall, and F1-score; Right: impact of $\alpha$ on IOU in terms of precision, recall, and F1-score. Dotted lines are results for SNMN, solid lines display results for SNMN-Caps. SNMN-Caps has significantly higher recall for all values of $\alpha$. However, scaling up opacity on uniform attention regions by $\alpha$ improves precision and consequently F1-score for SNMN-Caps. Results for SNMN are not effected by changing $\alpha$. Therefore, we choose $\alpha = 7$ to post process attention maps from both SNMN and SNMN-Caps to report final results.

inal SNMN model has clearly not learned this behavior and always focuses on some object in the last reasoning step which leads to false positive detections. To further investigate this subset of questions, we look into the capsules with highest probability for the last step before $No\_Op$ (no operation); we notice that capsule 0 was selected the most for the $Answer$ module (see figure 6). This validates our observation that capsules have learned to attend the background when no evidence is available for the answer. See figure 5 for attention visualizations with and without the capsule module.

Code and the CLEVR-Answers dataset will be released upon publication.

## References


[1] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 6

[2] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *International Conference on Learning Representations (ICLR)*, 2018. 1, 3