# Systematic Generalization on gSCAN:
# What is Nearly Solved and What is Next?

**Linlu Qiu** [1*]    **Hexiang Hu** [1 †]    **Bowen Zhang** [2]    **Peter Shaw** [1]    **Fei Sha** [1]

[1] Google Research        [2] USC

## Abstract

We analyze the *grounded* SCAN (gSCAN) benchmark, which was recently proposed to study systematic generalization for grounded language understanding. First, we study which aspects of the original benchmark can be solved by commonly used methods in multi-modal research. We find that a general-purpose Transformer-based model with cross-modal attention achieves strong performance on a majority of the gSCAN splits, surprisingly outperforming more specialized approaches from prior work. Furthermore, our analysis suggests that many of the remaining errors reveal the same fundamental challenge in systematic generalization of linguistic constructs regardless of visual context. Second, inspired by this finding, we propose challenging new tasks for gSCAN by generating data to incorporate relations between objects in the visual environment. Finally, we find that current models are surprisingly data inefficient given the narrow scope of commands in gSCAN, suggesting another challenge for future work.

## 1 Introduction

Systematic generalization refers to the ability to understand new compositions of previously observed concepts and linguistic constructs. While humans exhibit this ability, neural networks often struggle. To study systematic generalization, several synthetic datasets have been proposed. Lake and Baroni (2018) introduced SCAN, a dataset composed of natural language instructions paired with action sequences, and splits in various ways to assess systematic generalization. Recently, Ruis et al. (2020) introduced the grounded SCAN (gSCAN) benchmark, which similarly pairs natural language instructions with action sequences, but further requires that instructions are interpreted within the context of a grid-based visual navigation environment. In this work, we analyze which aspects of

gSCAN can currently be solved by general-purpose models, and propose new tasks and evaluation metrics for future research of systematic generalization on gSCAN.

First, to understand which aspects of gSCAN can be addressed by a general-purpose approach, we evaluate a Transformer-based model with cross-modal attention. Cross-modal attention has been proven effective for other multi-modal tasks (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020). It achieves strong performance on a majority of the splits, surprisingly outperforming several "specialist" approaches designed for gSCAN (Heinze-Deml and Bouchacourt, 2020; Gao et al., 2020; Kuo et al., 2020). We analyze the remaining errors, finding that many of the errors appear to be related to the same fundamental challenge in systematic generalization of linguistic constructs studied in datasets such as SCAN, regardless of the visual context.

Our analysis motivates the creation of an additional gSCAN task, which features a greater degree of complexity in how natural language instructions are grounded in the visual context. In this task, the agent needs to reason about spatial relations between objects expressed in language. We find this new task to be challenging for existing models.

Finally, we also assess the data efficiency of our cross-modal attention model on gSCAN. We find that despite the simplicity of the world state and the grammar used to generate instructions, model performance on most splits declines significantly when provided with less than ∼40% of the 360,000 original training examples. This suggests that we should consider sample complexity for future work.

## 2 Cross-modal Attention Solves gSCAN, Almost

**Experimental Setup** gSCAN has two types of generalization tasks: compositional generalization (CG) and length generalization. We focus on CG splits, which consist of a shared training set, 1 ran-

---

| Split | Seq2Seq (2020) | GECA (2020) | Kuo (2020) | Heinze (2020) | Gao (2020) | FiLM (2018) | RN (2017) | Ours |
|---|---|---|---|---|---|---|---|---|
| A | 97.69 ± 0.22 | 87.6 ± 1.19 | 96.73 ± 0.58 | 94.19 ± 0.71 | 98.60 ± 0.95 | 98.83 ± 0.32 | 97.38 ± 0.33 | **99.95 ± 0.02** |
| B | 54.96 ± 39.39 | 34.92 ± 39.30 | 94.91 ± 1.30 | 87.31 ± 4.38 | 99.08 ± 0.69 | 94.04 ± 7.41 | 49.44 ± 8.19 | **99.90 ± 0.06** |
| C | 23.51 ± 21.82 | 78.77 ± 6.63 | 67.72 ± 10.83 | 81.07 ± 10.12 | 80.31 ± 24.51 | 60.12 ± 8.81 | 19.92 ± 9.84 | **99.25 ± 0.91** |
| D | 0.00 ± 0.00 | 0.00 ± 0.00 | **11.52 ± 8.18** | - | 0.16 ± 0.12 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| E | 35.02 ± 2.35 | 33.19 ± 3.69 | 76.83 ± 2.32 | 52.8 ± 9.96 | 87.32 ± 27.38 | 31.64 ± 1.04 | 42.17 ± 6.22 | **99.02 ± 1.16** |
| F | 92.52 ± 6.75 | 85.99 ± 0.85 | 98.67 ± 0.05 | - | 99.33 ± 0.46 | 86.45 ± 6.67 | 96.59 ± 0.94 | **99.98 ± 0.01** |
| G | 0.00 ± 0.00 | - | **1.14 ± 0.30** | - | - | 0.04 ± 0.03 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| H | 22.70 ± 4.59 | 11.83 ± 0.31 | 20.98 ± 1.38 | - | **33.6 ± 20.81** | 11.71 ± 2.34 | 18.26 ± 1.24 | 22.16 ± 0.01 |

Table 1: Percentage of exact matches by various methods on gSCAN's compositional splits.

| Split | Held-out Examples |
|---|---|
| A: Random | Random |
| B: Yellow squares | *Yellow squares* referred with color and shape |
| C: Red squares | *Red squares* as target |
| D: Novel direction | Targets are located *south-west* of the agent |
| E: Relativity | Circles with size 2 and referred as *small* |
| F: Class inference | Squares with size 3 and need to be pushed twice |
| G: Adverb k = 1 | Commands with *cautiously* |
| H: Adverb to verb | Commands with *while spinning* and *pull* |

Table 2: Held-out examples on gSCAN's compositional splits.

dom test set, and 7 test sets of specifically designed held-out examples such that examples differ from the training set in various ways, cf. Table 2. The input commands (*e.g. walk to a red circle hesitantly*) are synthetically generated. The agent observes the world state which is a $d \times d$ grid ($d = 6$ in our case) with objects in various visual attributes. The output is an action sequence in the grid world (*e.g. turn left, walk, walk, stay*). We use exact match of the entire sequence as evaluation metric and report the mean and standard deviation across 5 runs.

**Our Model** We implement a seq2seq model with 6 Transformer layers for the encoder and the decoder each. The architecture is similar to ViL-BERT (Lu et al., 2019), a popular multi-modal model for visual and text information fusion. On a high-level, the encoder's text stream reads the commands while its visual stream encodes the world states. There are **cross-modal attention** between the two streams. The details are in Appendix A.

**Results** Table 1 shows the results by various models. Among them, FiLM and Relation Networks (RN) have achieved strong performance on other synthetic datasets for similar multi-modal tasks (Perez et al., 2018; Santoro et al., 2017). We also compare to models with specialized designs for CG or even gSCAN: a seq2seq model using data augmentation (Andreas, 2020), auxiliary loss (Heinze-Deml and Bouchacourt, 2020), task-specific architecture (Kuo et al., 2020; Gao



**Command**: pull a yellow big circle while spinning

**Prediction**: L_turn L_turn L_turn L_turn walk pull pull

**Target**: L_turn L_turn L_turn L_turn walk L_turn L_turn L_turn L_turn pull L_turn L_turn L_turn pull
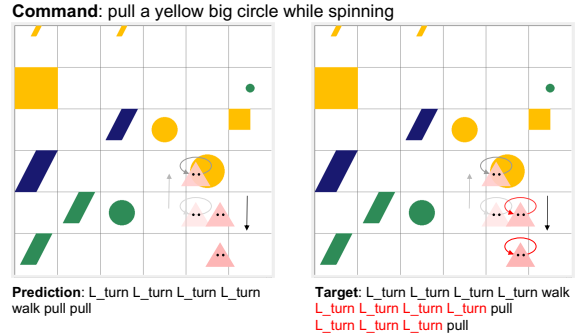
Figure 1: A failure case from the H split (Adverb to verb). Prediction (left) differs from target (right) in action sequences, while correct object is identified. Missing actions are in red.

et al., 2020).

The cross-modal attention model outperforms others on 5 out 8 splits. We hypothesize it is more effective as we leverage cross-modal attention to allow bi-directional interaction between language instruction and visual environment, in contrast to prior works that only have uni-directional attention from text to visual context (Ruis et al., 2020; Kuo et al., 2020; Gao et al., 2020; Heinze-Deml and Bouchacourt, 2020). The additional attention from visual context to text improves grounding natural language instructions to the visual environment. However, on the "hard" splits (D, G and H), all methods struggle.

**Analysis** First, we analyze in details the "hard" splits. Certain aspects of interpreting instructions are highly dependent on visual grounding. Most prominently, every instruction requires resolving the location of a referred object. However, adverbs such as *cautiously* or *while spinning* have the same meaning regardless of the visual context. Figure 1 shows one such example from the H split. The agent can successfully locate the target object but fails to combine the seen verb *pull* and adverb *while spinning* to generate the correct action sequence. Therefore, to assess the degree to which errors are caused by incorrect visual grounding, we calculate

| Split | Col Match↑ | Row Match↑ | Full Match↑ | No Match↓ |
|---|---|---|---|---|
| D | 32.91 ± 3.02 | 45.30 ± 6.46 | 0.00 ± 0.00 | 21.78 ± 3.58 |
| G | 98.43 ± 3.40 | 98.82 ± 2.52 | 97.91 ± 4.50 | 0.66 ± 1.42 |
| H | 91.23 ± 8.53 | 92.90 ± 6.54 | 88.16 ± 11.40 | 4.04 ± 3.71 |

Table 3: Percentage of exact matches of target position on splits the cross-modal attention model fails. For the novel direction split (D), the agent can end up at correct row or column around 80% of the time. For the adverb splits (G, H), the agent can usually find the object, but fails to generate action sequence with correct manner.



**Command:** pull a green small square hesitantly
**Prediction:** R_turn walk stay pull stay pull stay
**Target:** L_turn L_turn walk stay L_turn walk stay

**Command:** walk to a blue square
**Prediction:** R_turn walk
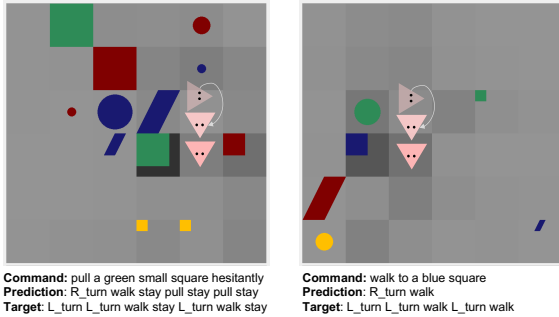**Target:** L_turn L_turn walk L_turn walk

Figure 2: Visualizations of predictions with attention weights for novel direction (D) split. The darker the grid, the higher the attention score.

the exact match between the target position and the agent's final position. The results are shown in Table 3. Despite the low exact match of entire sequences, the agent can find the correct targets more than 90% of time for two adverb splits (G, H). For the novel direction split (D), the direction "south west" is not seen during training. Similarly to Ruis et al. (2020), we find that the correct row *or* column is often selected, but not both. By analyzing attention weights, Ruis et al. (2020) attribute this as a failure to generate novel combinations of actions, not necessarily to identify the correct target object location, similar to our findings from visualizing attention weights in Figure 2. Therefore, we hypothesize that for these three splits (D, G, and H), the primary remaining challenge for the cross-modal attention model is not necessarily related to visual grounding.

Next, we analyze the splits (A, B, C, E, and F) where the cross-modal attention model performs well. Table 4 contrasts several variants of our model. We notice that removing decoder's textual attention (*i.e.*, −T) has virtual no significant change. On the other end, removing visual attention (−V) causes significantly worsened performance on E split. Removing cross-modal attention (−X) significantly degrades all splits, highlighting the benefit of cross-modal attention.

| Split | Ours | −V | −T | −X |
|---|---|---|---|---|
| A | 99.95 ± 0.02 | 99.73 ± 0.27 | **99.96 ± 0.01** | 96.69 ± 4.32 |
| B | **99.90 ± 0.06** | 97.55 ± 2.57 | 95.65 ± 5.48 | 31.42 ± 11.39 |
| C | **99.25 ± 0.91** | 97.71 ± 1.69 | 96.07 ± 4.61 | 32.39 ± 6.95 |
| E | **99.02 ± 1.16** | 87.67 ± 10.69 | 96.04 ± 5.15 | 43.19 ± 22.82 |
| F | **99.98 ± 0.01** | 99.85 ± 0.24 | **99.98 ± 0.01** | 96.50 ± 4.72 |

Table 4: Ablation studies of our model on gSCAN (-V, -T, and -X: removing decoder's visual, decoder's textual, and encoder's cross-modal attentions respectively).



**Training**

**Command:** walk to a circle *west of* a blue small square
**Target:** L_turn walk L_turn walk walk

**Testing**

**Command:** walk to a big circle *south west of* a green small cylinder
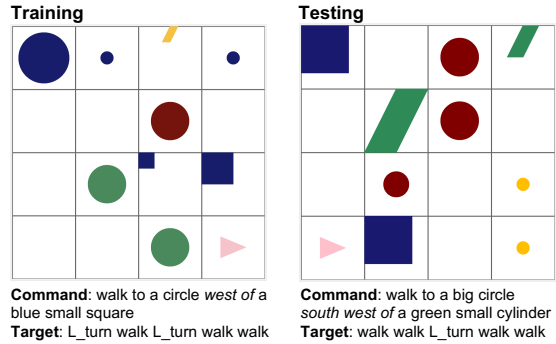**Target:** walk walk L_turn walk walk

Figure 3: Training and testing examples for spatial relation splits. The agent is shown as a pink triangle. The agent has seen relative position "west" (left) and "south", "east", "south east" (in other examples) during training and needs to generalize to "south west" (right) during testing.

# 3 Grounded Spatial Relation CG

**Proposed Task** Given our analysis suggesting that many of the remaining challenges for gSCAN may not necessarily be related to visual grounding, we propose an additional task that features a greater degree of complexity in how natural language instructions are grounded in the visual environment. We hope this will complement the original gSCAN tasks as a useful assessment of systematic generalization in grounded language understanding.

The new data we create will contain language expressions that refer to target objects along with their relations to a second referenced object. We use two types of relations: *next to* and relative positions such as *north* and *west*. As an example, we can have expressions such as *a blue square next to a red circle* or *a blue square north of a red circle*.

To ensure that interpreting the spatial relation is necessary to correctly identify the target object, we create *visual distractors* in the environment. In the above examples, a blue square that is north of a magenta circle would be a distractor to the intended blue square which is next to the red circle. The existence of visual distractors will force the agent to examine the correspondence between the visual information and the (compositional) lan-

| Split | Held-out Examples |
|---|---|
| I: Random | Random |
| II: Visual | *Red squares* as target or reference |
| III: Relation | *Green squares* and *blue circles* combinations |
| IV: Referent | *Yellow squares* referred as target |
| V: Relative position 1 | Targets are *north* of their references |
| VI: Relative position 2 | Targets are *south west* of their references |

Table 5: Held-out examples for spatial relation learning task.

| Split | Seq2Seq (2020) | FiLM (2018) | RN (2017) | Ours |
|---|---|---|---|---|
| I | $86.48 \pm 0.64$ | $88.85 \pm 0.82$ | $85.17 \pm 3.81$ | $\mathbf{94.66} \pm \mathbf{0.24}$ |
| II | $40.10 \pm 0.83$ | $50.68 \pm 0.32$ | $38.59 \pm 0.74$ | $\mathbf{64.41} \pm \mathbf{4.52}$ |
| III | $86.08 \pm 0.73$ | $88.81 \pm 1.42$ | $85.66 \pm 4.35$ | $\mathbf{94.89} \pm \mathbf{0.20}$ |
| IV | $5.47 \pm 0.09$ | $10.78 \pm 3.47$ | $4.85 \pm 0.86$ | $\mathbf{49.58} \pm \mathbf{3.47}$ |
| V | $81.41 \pm 1.03$ | $76.20 \pm 2.64$ | $79.86 \pm 3.16$ | $59.29 \pm 5.63$ |
| VI | $81.84 \pm 1.38$ | $75.05 \pm 3.63$ | $80.93 \pm 2.76$ | $49.50 \pm 6.49$ |

Table 6: Percentage of exact matches by various methods on proposed spatial relation splits.

| Split | Ours | $-$V | $-$T | $-$X |
|---|---|---|---|---|
| I | $94.66 \pm 0.24$ | $\mathbf{94.70} \pm \mathbf{0.23}$ | $94.58 \pm 0.26$ | $92.71 \pm 0.90$ |
| II | $64.41 \pm 4.52$ | $62.42 \pm 7.12$ | $59.75 \pm 4.18$ | $46.65 \pm 2.20$ |
| III | $94.89 \pm 0.20$ | $94.90 \pm 0.25$ | $\mathbf{94.91} \pm \mathbf{0.30}$ | $92.88 \pm 0.79$ |
| IV | $49.58 \pm 3.47$ | $36.15 \pm 16.42$ | $35.07 \pm 13.73$ | $10.76 \pm 1.76$ |
| V | $59.29 \pm 5.63$ | $58.95 \pm 6.48$ | $59.21 \pm 6.39$ | $\mathbf{86.34} \pm \mathbf{1.50}$ |
| VI | $49.50 \pm 6.49$ | $49.40 \pm 5.66$ | $47.91 \pm 8.28$ | $\mathbf{87.10} \pm \mathbf{2.15}$ |

Table 7: Ablation studies of our model on new relation reasoning task (-V, -T, and -X: removing decoder's visual, decoder's textual, and encoder's cross-modal attentions respectively).

guage expression to disambiguate and locate the correct target object.

Similar to the gSCAN setup, we use a shared training set and hold-out specific examples to evaluate generalization abilities for learning the visually grounded spatial relation reasoning as in Figure 3. In addition to random split (I: Random), we create new test splits including novel object properties (II: Visual), novel target and reference combinations (III: Relation), novel referent (IV: Referent), and novel relative positions (V, VI: Relative position).

Table 5 shows detailed descriptions of the held-out examples for each test split. In total, we generated about 260,000 examples for the new task, using 18 templates and about 32,000 unique instructions.[1] Details are included in Appendix B.

**Results** We evaluate the cross-modal attention model on this new task and show results in Table 6. The model outperforms the baseline methods by a large margin on 4 out of 6 splits, but performs surprisingly worse on splits V and VI. The model performs unexpectedly well on the III. We conduct ablation studies and report the results in Table 7. III is also robust to various ways of removing attentions, similar to the F split we observed in Table 4. On the other end, the splits II and IV are affected negatively very much, while V and VI are surprisingly improved with removing the cross-modal attention. However, there is not a single model that excels at all splits at the same time.

We hypothesize that the cross-modal attention model overfits to certain aspects of the training distribution, leading to worse out of distribution performance on V and VI splits. We compute the exact match of our models on examples of seen relations from the random test split, to verify the in-domain generalization. Particularly, this in-domain settings evaluate situations where targets are "north west/north east" (in-domain V) and "south/west" (in-domain VI) to their references. It

shows that the cross-modal attention model (95.52 on in-domain V and 95.43 on in-domain VI) has better in-domain generalization than its without cross-modal counterpart (93.14 on in-domian V and 93.90 on in-domain VI), but it performs worse in compositional generalization. Regularization techniques could potentially be used to improve the cross-modal attention model, which we leave for future research.

## 4 Sample Complexity

One inspiration of studying systematic generalization is to develop techniques that can reduce the sample complexity for learning novel behaviors. While we observe that many models perform well at least on some splits, we ask a natural question that has not been studied before: *even on those splits, are our models fulfilling the promise of learning systematic generalization?*

Figure 4 shows that for the original gSCAN and the newly proposed task, performance for the cross-modal attention model starts to significantly drop when trained using less than around 40% of the training data for most splits. The model without cross-modal attention is even more data inefficient; performance starts to significantly drop when trained using less than 70% of the training data for the original compositional splits, or when reducing the training data by any amount for the spatial relation splits. This suggests exploring model architectural priors can help improve data efficiency and provide more benefits for generalization.
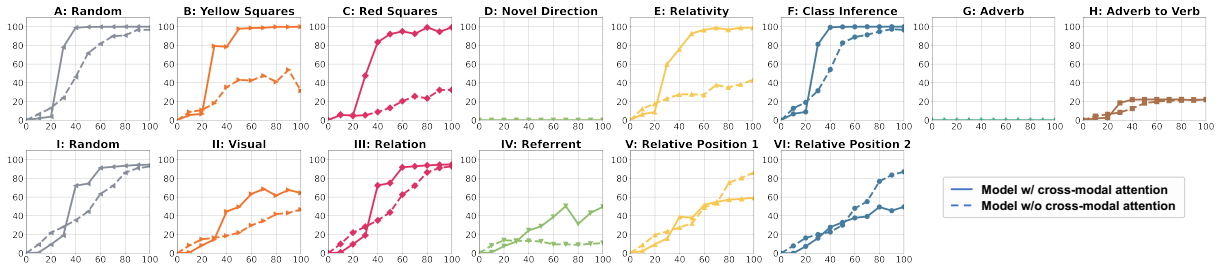
Figure 4: Data efficiency of the models on compositional splits (top) and spatial relation splits (bottom). The x-axis is the percentage of data (%) and the y-axis is the exact match (%).

To investigate how the number of primitives influences the data efficiency, we re-generate smaller compositional splits and spatial relation splits by reducing the number of primitives. For compositional splits, we exclude one noun (cylinder), one color adjective (blue), and one adverb (while zigzagging). For spatial relation splits, we exclude one noun (cylinder) and one location preposition (next to). This reduces the number of training examples by around 2/3 and leads to around 110,000 and 74,000 training examples for the smaller compositional splits and spatial relation splits respectively.
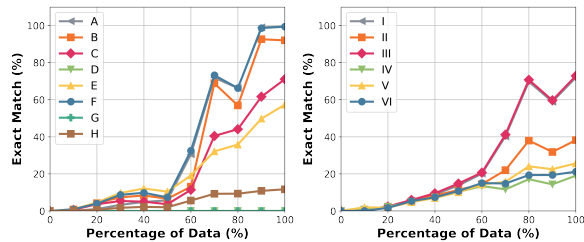


Figure 5: Data efficiency of the model with cross-modal attention on compositional splits (left) and spatial relation splits (right) with smaller number of primitives.

We then perform similar experiments with our cross-modal attention model. The results are shown in Figure 5. When the number of primitives decreases, the percentage of training data required to attain satisfactory performance is increased, with more than 80% of training data needed for both splits. One possible explanation is that the total number of training examples is reduced when the number of primitives decreases. Therefore, the model sees fewer combinations during training, thus needs a higher percentage of data to properly learn visual grounding and compositionality. This provides further evidence that current model need extensive training data to achieve systematic generalization, demonstrating the necessity of evaluating sample complexity for future work.

## 5 Related work

Many datasets and tasks have been proposed for examining systematic generalization: visual question answering (Johnson et al., 2017; Bahdanau et al., 2019a; Pezzelle and Fernández, 2019; Bahdanau et al., 2019b), visually grounded navigation instruction following (Hermann et al., 2017; Yu et al., 2018; Chevalier-Boisvert et al., 2019; Chaplot et al., 2018). Many approaches have been proposed too, though a large percentage of them uses task-specific design and only work well for the specific tasks or datasets, for example, learning neural program executions (Andreas et al., 2016; Hu et al., 2017; Johnson et al., 2017; Santoro et al., 2017; Hudson and Manning, 2018; Mao et al., 2019), in addition to the ones we have described previously.

This work focuses on using a generic cross-modal attention model to probe the gSCAN dataset in the hope of understanding in what aspect the task/dataset is challenging. Ding et al. (2020) also presented new evidences that a neural-based model can solve similar CG tasks. Our interest is to use such models to inspire new task designs that continue to challenge neural models.

## 6 Conclusion

In this work, we have demonstrated that a general-purpose cross-modal attention model can achieve strong performance on a majority of gSCAN splits and outperform more specialized prior work. We have proposed a challenging additional task for gSCAN that requires agents to reason over spatial relations between objects in the visual scene, and have highlighted data efficiency as a consideration for future work.

## Acknowledgements

# References

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society.

Dzmitry Bahdanau, Harm de Vriesand Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. 2019a. CLOSURE: assessing systematic generalization of CLEVR models. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop*.

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron C. Courville. 2019b. Systematic generalization: What is required and can it be learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. 2018. Gated-attention architectures for task-oriented language grounding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2819–2826. AAAI Press.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. Babyai: A platform to study the sample efficiency of grounded language learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

David Ding, Felix Hill, Adam Santoro, and Matt Botvinick. 2020. Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. *ArXiv preprint*, abs/2012.08508.

Tong Gao, Qi Huang, and Raymond Mooney. 2020. Systematic generalization on gSCAN with language conditioned embedding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 491–503, Suzhou, China. Association for Computational Linguistics.

Christina Heinze-Deml and Diane Bouchacourt. 2020. Think before you act: A simple baseline for compositional generalization. *ArXiv preprint*, abs/2009.13962.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. 2017. Grounded language learning in a simulated 3d world. *ArXiv preprint*, abs/1706.06551.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. 2017. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 804–813. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Yen-Ling Kuo, Boris Katz, and Andrei Barbu. 2020. Compositional networks enable systematic generalization for grounded language understanding. *ArXiv preprint*, abs/2008.02742.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press.

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.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. 2020. A benchmark for systematic generalization in grounded language understanding.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4967–4976.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Haonan Yu, Haichao Zhang, and Wei Xu. 2018. Interactive grounded language acquisition and generalization in a 2d world. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

# Appendix

In this appendix, we include more details about our cross-modal attention model (§ A) and data generation procedure (§ B).

## A  Models

In this section, we include more details about our cross-modal attention model to ensure all results are fully reproducible. We also include details of the baseline methods we implement.

### A.1  Model details

We show our model architecture in Figure 6. The model follows the standard encoder-decoder structure in Transformer (Vaswani et al., 2017) and uses cross-modal attention in the encoder similar to ViLBERT (Lu et al., 2019). It is trained via teacher forcing. During test time, it generates output actions in the auto-regressive manner.

**Encoder**  The model takes input command tokens $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ and world state $\mathbf{W} \in \mathbb{R}^{d \times d \times c}$, where $d$ is the grid size and $c$ is world state dimension as in the baseline seq2seq model (Ruis et al., 2020). The encoder maps the world state to visual representation $\mathbf{H}^v$ through multi-scale convolutional networks followed by linear layers. It encodes the input tokens to input embeddings $\mathbf{H}^l = \{\mathbf{h}_1^l, \mathbf{h}_2^l, ..., \mathbf{h}_n^l\}$ with positional encoding.

The visual and linguistic representations are passed through $N = 6$ transformer blocks with cross-modal attention. Similar to ViLBERT (Lu et al., 2019), each transformer block consists of two parallel multi-head attention blocks, with the representation of one modality passed as key and value to the multi-head attention block of the other modality.

**Decoder**  The decoder consists of $N = 6$ stacked blocks similar to the decoder in Transformer (Vaswani et al., 2017). Each block contains one self-attention block and one multi-head attention block over the contextual embeddings $\mathbf{H}^c = [\mathbf{H}^v; \mathbf{H}^l]$ of the encoder.

### A.2  Implementation Details

We use $N = 6$ layers of transformer blocks for the encoder and the decoder. Each block has hidden size of 128, intermediate size of 256, and 8 attention heads. The total number of parameters of the model is around 3M for all model variants. We train our model for 40 epochs with batch size of 128. We use the Adam optimizer with weight decay. The initial learning rate is set to be 0.0016 for model with cross-modal attention and 0.0008 for model without it. We use linear warmup for the first 10% of training epoch and decrease the learning rate by 10 in the 20th and 30th epoch. The dev set accuracy and exact match are used to validate model performance.
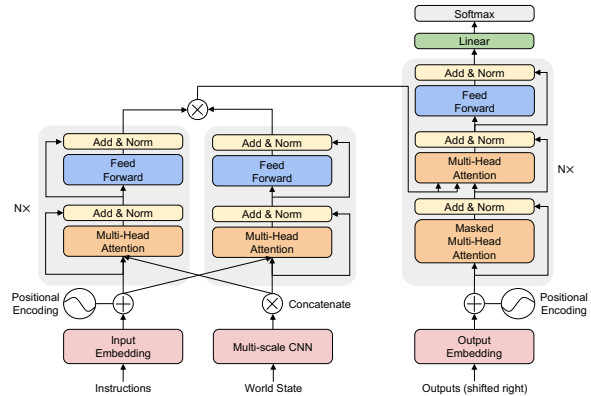


Figure 6: The architecture of our model. The encoder consists of 6 layers of transformer block with cross-modal attention. The decoder contains 6 layers of transformer block, and each block has one self-attention block and one multi-head attention block over contextual representations of encoder. [2]

### A.3  Baseline Methods

We implement FiLM (Perez et al., 2018) and RN (Santoro et al., 2017) as our baselines. The baselines are built upon the seq2seq model. FiLM learns functions to predict a set of $\beta$, $\gamma$ conditioned on inputs to modulate neural networks. For FiLM, we add a linear layer to predict $\beta$, $\gamma$ for each convolution layer. For RN, we use 3 relation layers with hidden size of 256 followed by 2 fully-connected layers to predict object relations. The hidden states are then used to initialize the LSTM decoder. We refer the readers to the original FiLM and RN paper for model details. Other hyper-parameters remain the same as the original seq2seq baseline.[3]

## B  Data Generation

We build upon the original gSCAN codebase to generate spatial relation data splits.[4]

---

[2]The figure is modified from https://github.com/dair-ai/ml-visuals.

[3]https://github.com/LauraRuis/multimodal_seq2seq_gSCAN.

[4]https://github.com/LauraRuis/groundedSCAN.

## B.1 Input Commands Generation

The input commands are generated based on context-free grammar (CFG) in Table 8. We add new grammar rules and lexicon rules to support object relations. Since we focus on evaluating relation reasoning, we exclude templates without any object relation.

| Grammar Rule |
|---|
| ROOT → VP |
| VP → VV$_i$ 'to' DP |
| VP → VV$_t$ DP |
| DP → 'a' NP |
| NP → JJ NP |
| NP → NP PP |
| PP → LOC DP |
| NP → NN |
| **Lexicon Rule** |
| VV$_i$ → {walk} |
| VV$_t$ → {push, pull} |
| NN → {circle, square, cylinder} |
| JJ → {red, green, blue, yellow, big, small} |
| LOC → {next to} |
| LOC → {east of, north of, west of, south of} |
| LOC → {north east of, north west of} |
| LOC → {south east of, south west of} |

Table 8: The CFG used to generate input commands.

## B.2 World State Generation

We generate the world state with the following constraints: 1) For each target referent $\mathbf{T}$, there will be one unique reference object $\mathbf{R}$ next to it; 2) There will be up to $n$ visual distractors $\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_n$, which have the same size, shape, and color of the target object $\mathbf{T}$; 3) Each visual distractor $\mathbf{V}_i$ will or will not have its own reference object $\mathbf{O}_i$.

Additionally, to avoid ambiguity, when generating visual distractor $\mathbf{V}_i$, we ensure: 1) If the input command contains abstract relative position (*i.e. next*), $\mathbf{V}_i$ cannot be placed near $\mathbf{R}$, and $\mathbf{O}_i$ must be distinct from $\mathbf{R}$; 2) If the input command contains specific relative position (*e.g. north, west*, etc.), $\mathbf{V}_i$ can be placed anywhere, and $\mathbf{O}_i$ can be the same as $\mathbf{R}$. However, $\mathbf{O}_i$ cannot have the same relative position to $\mathbf{V}_i$ as $\mathbf{R}$ to $\mathbf{T}$. Other procedures remain the same as the original setup.

## B.3 Data Examples

We show data examples for each spatial relation split in Figure 7. The systematic difference between the training and test split is highlighted.
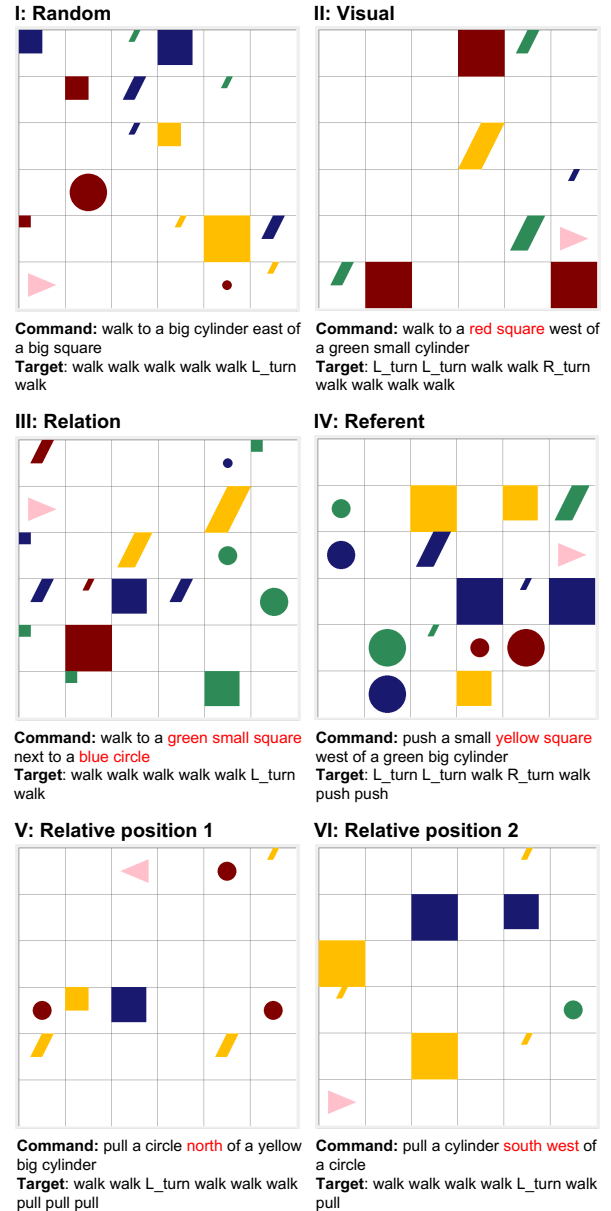


Figure 7: Data examples for each spatial relation test split with the systematic difference highlighted.