

# Syntax-Guided Transformers: Elevating Compositional Generalization and Grounding in Multimodal Environments

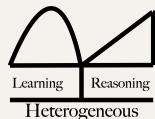
**Danial Kamali, Parisa Kordjamshidi**

Department of Computer Science and Engineering, Michigan State University

*kamalida@msu.edu, kordjams@msu.edu*



MICHIGAN STATE  
UNIVERSITY



---

# Compositional Generalization

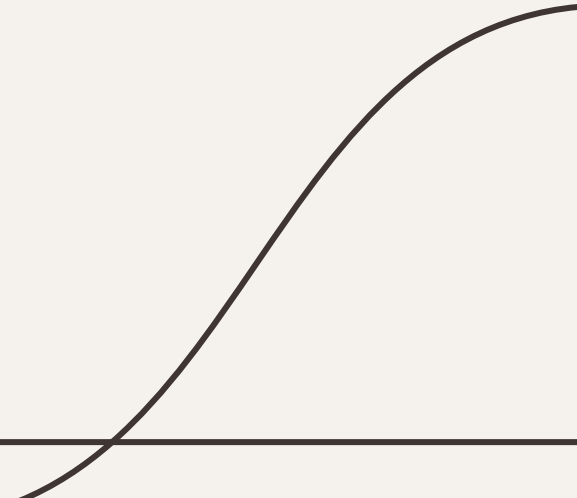
*The ability of intelligent models to  
extrapolate understanding of  
components to novel compositions*

---

---

# Grounded Language Understanding & Compositional Generalization

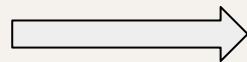
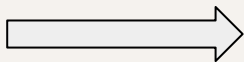
**Comprehend** and **apply language commands** in a multimodal setting while focusing on compositional generalization capabilities, such as **understanding** and **combining** known **words** and **concepts** in **novel** ways.



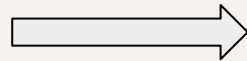
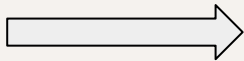
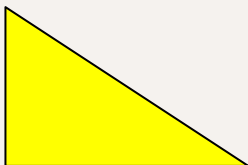
---

# Compositional Generalization

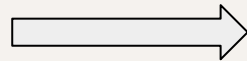
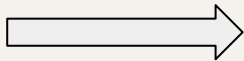
## Training



Yellow Rectangle



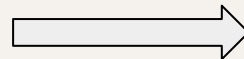
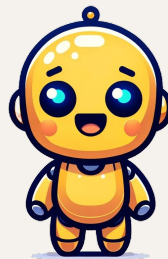
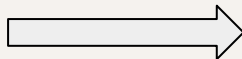
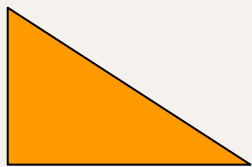
Yellow Triangle



Orange Rectangle

# Compositional Generalization

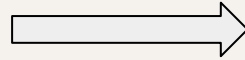
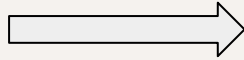
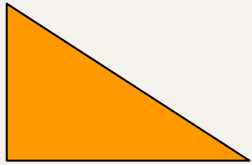
Testing



?

# Compositional Generalization

Testing



Orange Rectangle

# Grounded Language Understanding

An agent is provided with a command. Its objective is to generate/execute a series of predefined actions to fulfill the task within the given environment

- **Benchmarks**

- gSCAN
- GRRS
- ReaSCAN

- **Input**

- 6\*6 grid
- Input Command

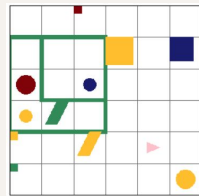
- **Output:**

- Actions: {turn\_left, turn\_right, walk, push, pull}

- **Evaluation Metric**

- Exact Match(%)

*pull the small blue object that is inside of the small green box and in the same row as the red circle while zigzagging*



*turn left,  
turn left,  
walk,  
turn right,  
walk,  
turn right,  
walk,  
pull*

# Grounded Language Understanding Challenges

- New composition of relative information
  - B2
    - **Test:** [obj desc] that is **same size as** [obj desc] and **inside of** [obj desc]
- Higher depth of reasoning
  - C1
    - **Train:** [obj desc] that is [rel] [obj desc] and [rel] [obj desc]
    - **Test:** [obj desc] that is [rel] [obj desc] and [obj desc] **and [rel] [obj desc]**
  - C2
    - **Train:** [obj desc] that is [rel] [obj desc] and [rel] [obj desc]
    - **Test:** [obj desc] that is [rel] [obj desc] **that is** [rel] [obj desc]

Split	Held-out Examples
Random	Random.
A1	<i>yellow square</i> referred with color & shape.
A2	<i>red square</i> referred in the command.
A3	<i>small cylinder</i> referred with size and shape
B1	co-occur of <i>small red circle</i> and <i>big blue square</i> .
B2	co-occur of <i>same size as</i> and <i>inside of</i> relations.
C1	Additional conjunction clause depth added to <i>2-relative-clause commands</i> .
C2	<i>2-relative-clause</i> command with <i>that is</i> instead of <i>and</i> .

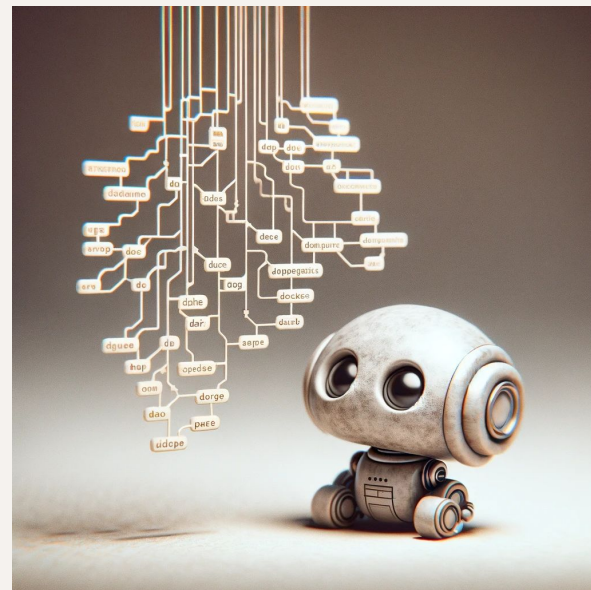
Table 1: ReaSCAN dataset test splits.

Model	A1	A2	A3	B1	B2	C1	C2	Avg
LSTM*	50.4	14.7	50.9	52.2	39.4	49.7	25.7	40.40
GCN-LSTM	92.3	42.1	87.5	69.7	52.8	57.0	22.1	60.50
Transformer*	96.7	58.9	93.3	79.8	59.3	75.9	25.5	69.90
GroCoT	99.6	93.1	98.9	93.9	86.0	76.3	27.3	82.2

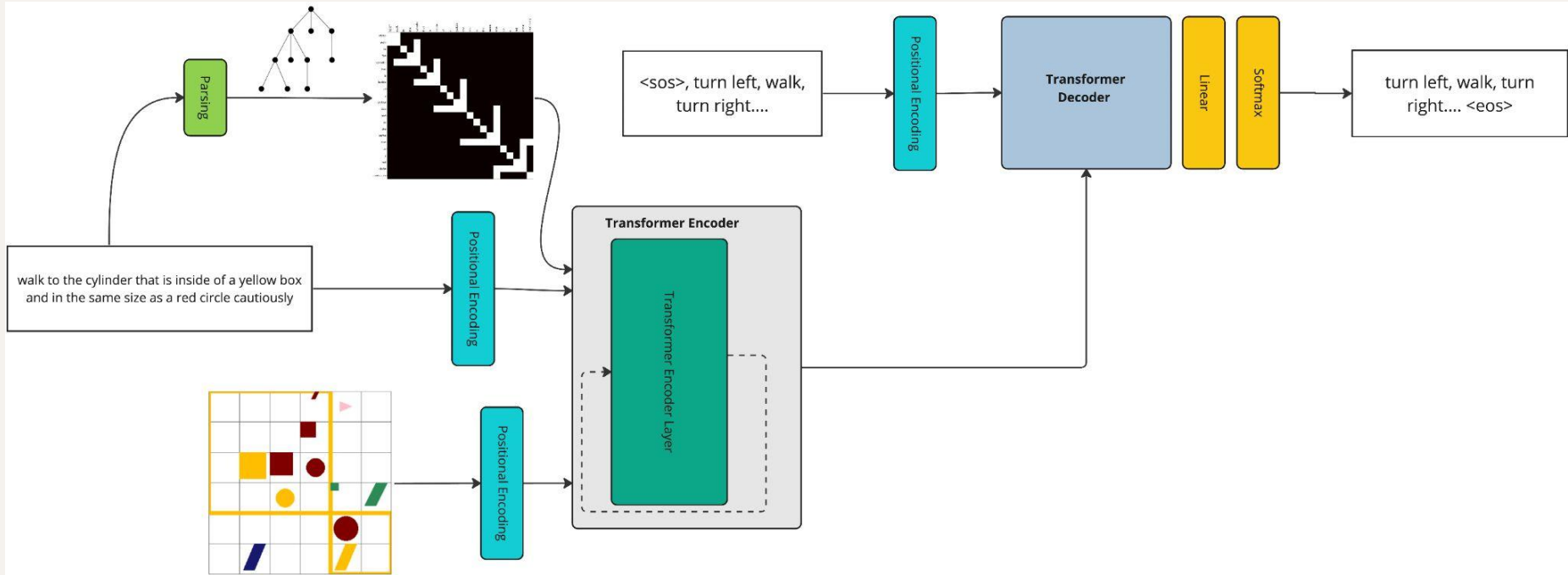


# Motivation

- **Syntactic Structure as a Key to Generalization:**
  - Utilizing readily available parsers to infer hints about the underlying syntactic structure.
  - Removing self-attention connection instead of adding complexity
- **Efficacy With Weight Sharing:**
  - Addressing the backpropagation challenges in attention masking methods through weight sharing.
  - Enhancing efficiency in model performance.



# Model



# Results

Model	A1	A2	A3	B1	B2	C1	C2	Avg
LSTM*	50.4	14.7	50.9	52.2	39.4	49.7	25.7	40.40
GCN-LSTM	92.3	42.1	87.5	69.7	52.8	57.0	22.1	60.50
Transformer*	96.7	58.9	93.3	79.8	59.3	75.9	25.5	69.90
GroCoT	99.6	93.1	98.9	93.9	86.0	76.3	<b>27.3</b>	82.2
Constituency <sup>†</sup>	<b>99.75</b> $\pm$ 0.11	96.70 $\pm$ 1.40	<b>99.68</b> $\pm$ 0.10	95.19 $\pm$ 1.17	88.37 $\pm$ 1.50	69.07 $\pm$ 0.60	27.00 $\pm$ 0.54	82.25 $\pm$ 0.63
Dependency <sup>†</sup>	99.65 $\pm$ 0.9	<b>97.37</b> $\pm$ 0.48	99.62 $\pm$ 0.07	<b>95.46</b> $\pm$ 2.01	<b>90.15</b> $\pm$ 3.88	<b>92.55</b> $\pm$ 1.51	21.77 $\pm$ 5.25	<b>85.22</b> $\pm$ 0.87

Table 2: The result of our proposed model on the ReaSCAN dataset test splits. The results are an average of three runs. † denotes the models with masking. Models marked with \* refer to the multimodal version of their implementation.

# Extended Results

Model	A	B	C	E	F	H	Comp. Avg
LSTM*	97.7	54.9	23.5	35.0	92.5	22.7	32.7
GCN-LSTM	98.6	99.1	80.3	87.3	99.3	<b>33.6</b>	-
Transformer*	99.9	99.9	99.3	99.0	99.9	22.2	60.0
GroCoT	99.9	99.9	99.9	99.8	99.9	22.9	60.4
Constituency <sup>†</sup>	<b>99.95</b> $\pm$ 0.07	<b>99.92</b> $\pm$ 0.06	<b>99.88</b> $\pm$ 0.11	99.88 $\pm$ 0.09	<b>100.00</b> $\pm$ 0.00	22.84 $\pm$ 0.93	60.36 $\pm$ 0.11
Dependency <sup>†</sup>	99.92 $\pm$ 0.09	99.85 $\pm$ 0.18	99.86 $\pm$ 0.11	<b>99.96</b> $\pm$ 0.06	99.89 $\pm$ 0.16	23.89 $\pm$ 1.54	<b>60.49</b> $\pm$ 0.20

Table 3: The result of our proposed model on the gSCAN dataset test splits. The results are an average of three runs. We did not report the results on D and G splits since we achieved 0.00 $\pm$ 0.00 % performance, But take them into account in the averaged result. <sup>†</sup> denotes the models with masking. Models marked with \* refer to the multimodal version of their implementation.

Model	I	II	III	IV	V	VI	Comp. Avg
LSTM*	86.5	40.1	86.1	5.5	81.4	81.8	58.9
Transformer*	94.7	64.4	94.9	49.6	59.3	49.5	63.5
GroCoT	99.9	98.6	<b>99.9</b>	99.7	<b>99.5</b>	96.5	98.8
Constituency <sup>†</sup>	99.85 $\pm$ 0.00	99.90 $\pm$ 0.03	99.16 $\pm$ 0.26	99.88 $\pm$ 0.03	96.73 $\pm$ 2.16	<b>97.85</b> $\pm$ 0.46	98.58 $\pm$ 0.39
Dependency <sup>†</sup>	<b>99.91</b> $\pm$ 0.02	<b>99.93</b> $\pm$ 0.01	99.41 $\pm$ 0.28	<b>99.96</b> $\pm$ 0.01	99.03 $\pm$ 0.23	97.38 $\pm$ 0.63	<b>99.07</b> $\pm$ 0.16

Table 4: The result of our proposed model on the GSRR dataset test splits. The results are an average of three runs. <sup>†</sup> denotes the models with masking. Models marked with \* refer to the multimodal version of their implementation.

# Analysis

- Ablation Study

W/S	Mask	A1	A2	A3	B1	B2	C1	C2	Avg
-	-	99.29 $\pm$ 0.27	91.82 $\pm$ 6.50	98.49 $\pm$ 1.17	93.50 $\pm$ 0.85	83.15 $\pm$ 1.41	75.85 $\pm$ 1.35	<b>25.03</b> $\pm$ 6.82	81.02 $\pm$ 0.22
✓	-	99.68 $\pm$ 0.22	97.09 $\pm$ 1.72	99.64 $\pm$ 0.20	94.86 $\pm$ 0.77	81.49 $\pm$ 4.27	66.30 $\pm$ 6.65	21.66 $\pm$ 1.83	80.10 $\pm$ 1.08
-	Dep.	98.09 $\pm$ 0.27	85.21 $\pm$ 6.85	97.35 $\pm$ 0.75	93.61 $\pm$ 2.75	90.62 $\pm$ 1.59	75.27 $\pm$ 1.77	21.91 $\pm$ 1.63	80.29 $\pm$ 1.43
✓	Dep.	<b>99.65</b> $\pm$ 0.9	<b>97.37</b> $\pm$ 0.48	<b>99.62</b> $\pm$ 0.07	<b>95.46</b> $\pm$ 2.01	<b>90.15</b> $\pm$ 3.88	<b>92.55</b> $\pm$ 1.51	21.77 $\pm$ 5.25	<b>85.22</b> $\pm$ 0.87

Table 5: The ablation study result of our modifications on ReaSCAN dataset test splits. Results are reported on an average of three runs. We evaluate every combination of components from our best model. W/S stands for weight sharing, and the ✓ shows the presence of the module. *Dep* in this table refers to the Dependency masking. We evaluate the model with or without dependency masking in the masking part.

- Efficacy Analysis

Model	#Parameters
Multimodal LSTM	74K
Multimodal Transformer	3M
GroCoT	4.6M
Dependency <sup>†</sup> (ours)	1.9M

Table 6: Comparing model parameters: our model vs. current state-of-the-art models. Dependency<sup>†</sup> refers to the model with dependency parsing for attention masking.

---

# Conclusion

- Exploiting syntactic structure with weight sharing in Transformer encoders significantly improves generalization.
  - Using Dependency parsing was more effective than constituency parsing.
  - Using weight sharing with dependency parsing alleviates the backpropagation problem caused by attention masking.
-

---

# Thanks!

kamalida@msu.edu

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

