Supplementary Material for Abstract Spatial-Temporal Reasoning via Probabilistic Abduction and Execution

Chi Zhang⋆ Baoxiong Jia⋆ Song-Chun Zhu Yixin Zhu
UCLA Center for Vision, Cognition, Learning, and Autonomy
{chi.zhang,baoxiongjia}@ucla.edu, sczhu@stat.ucla.edu, yixin.zhu@ucla.edu

1. Step-by-Step Example

Here, we give a concrete example of the computation process on the panel attribute of `Number`. `Number` is dependent on the object attribute of `Objectiveness`. Assume the object CNN produces 4 objectiveness distributions:

$$\begin{aligned} P(\texttt{Obj}_1) &= [0.1, 0.9] \\ P(\texttt{Obj}_2) &= [0.7, 0.3] \\ P(\texttt{Obj}_3) &= [0.9, 0.1] \\ P(\texttt{Obj}_4) &= [0.8, 0.2], \end{aligned} \tag{1}$$

where the first element corresponds to the objectiveness being false.

Then the scene inference engine works out the `Number` distribution as

$$\begin{aligned} P(\texttt{Num} = 1) = {} & P(\texttt{Obj}_1 = \text{True}) \times P(\texttt{Obj}_2 = \text{False}) \\ & \times P(\texttt{Obj}_3 = \text{False}) \times P(\texttt{Obj}_4 = \text{False}) \\ & + P(\texttt{Obj}_1 = \text{False}) \times P(\texttt{Obj}_2 = \text{True}) \\ & \times P(\texttt{Obj}_3 = \text{False}) \times P(\texttt{Obj}_4 = \text{False}) \\ & + P(\texttt{Obj}_1 = \text{False}) \times P(\texttt{Obj}_2 = \text{False}) \\ & \times P(\texttt{Obj}_3 = \text{True}) \times P(\texttt{Obj}_4 = \text{False}) \\ & + P(\texttt{Obj}_1 = \text{False}) \times P(\texttt{Obj}_2 = \text{False}) \\ & \times P(\texttt{Obj}_3 = \text{False}) \times P(\texttt{Obj}_4 = \text{True}) \\ P(\texttt{Num} = 2) = {} & \dots \\ P(\texttt{Num} = 3) = {} & \dots \\ P(\texttt{Num} = 4) = {} & P(\texttt{Obj}_1 = \text{True}) \times P(\texttt{Obj}_2 = \text{True}) \\ & \times P(\texttt{Obj}_3 = \text{True}) \times P(\texttt{Obj}_4 = \text{True}). \end{aligned} \tag{2}$$

After normalization, we have

$$P(\texttt{Num}) = [0.51, 0.39, 0.09, 0.01], \tag{3}$$

where each element corresponds to `Number` being 1, 2, 3, and 4, respectively.

For each panel, we have one such distribution of `Number`. Assume `Number` distributions for 8 context panels are:

$$\begin{aligned} P(\texttt{Num}_1) &= [0.51, 0.39, 0.09, 0.01] \\ P(\texttt{Num}_2) &= [0.39, 0.51, 0.09, 0.01] \\ P(\texttt{Num}_3) &= [0.39, 0.09, 0.51, 0.01] \\ P(\texttt{Num}_4) &= [0.51, 0.39, 0.09, 0.01] \\ P(\texttt{Num}_5) &= [0.09, 0.01, 0.51, 0.39] \\ P(\texttt{Num}_6) &= [0.39, 0.09, 0.01, 0.51] \\ P(\texttt{Num}_7) &= [0.51, 0.39, 0.09, 0.01] \\ P(\texttt{Num}_8) &= [0.09, 0.51, 0.39, 0.01]. \end{aligned} \tag{4}$$

The probabilistic abduction engine computes each rule probability by

$$\begin{aligned} P(r^{\texttt{Num}} = \texttt{Const}) = {} & P(\texttt{Num}_1 = 1) \times P(\texttt{Num}_2 = 1) \times P(\texttt{Num}_3 = 1) \\ & \times P(\texttt{Num}_4 = 1) \times P(\texttt{Num}_5 = 1) \times P(\texttt{Num}_6 = 1) \\ & \times P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 1) \\ & + P(\texttt{Num}_1 = 2) \times P(\texttt{Num}_2 = 2) \times P(\texttt{Num}_3 = 2) \\ & \times P(\texttt{Num}_4 = 1) \times P(\texttt{Num}_5 = 1) \times P(\texttt{Num}_6 = 1) \\ & \times P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 1) \\ & \dots \\ & + P(\texttt{Num}_1 = 4) \times P(\texttt{Num}_2 = 4) \times P(\texttt{Num}_3 = 4) \\ & \times P(\texttt{Num}_4 = 4) \times P(\texttt{Num}_5 = 4) \times P(\texttt{Num}_6 = 4) \\ & \times P(\texttt{Num}_7 = 4) \times P(\texttt{Num}_8 = 4) \\ & \vdots \\ P(r^{\texttt{Num}} = \texttt{Plus}) = {} & P(\texttt{Num}_1 = 1) \times P(\texttt{Num}_2 = 1) \times P(\texttt{Num}_3 = 2) \\ & \times P(\texttt{Num}_4 = 1) \times P(\texttt{Num}_5 = 1) \times P(\texttt{Num}_6 = 2) \\ & \times P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 1) \\ & + P(\texttt{Num}_1 = 2) \times P(\texttt{Num}_2 = 1) \times P(\texttt{Num}_3 = 3) \\ & \times P(\texttt{Num}_4 = 1) \times P(\texttt{Num}_5 = 1) \times P(\texttt{Num}_6 = 2) \\ & \times P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 1) \\ & \dots \\ & + P(\texttt{Num}_1 = 3) \times P(\texttt{Num}_2 = 1) \times P(\texttt{Num}_3 = 4) \\ & \times P(\texttt{Num}_4 = 3) \times P(\texttt{Num}_5 = 1) \times P(\texttt{Num}_6 = 4) \\ & \times P(\texttt{Num}_7 = 3) \times P(\texttt{Num}_8 = 1) \\ & \vdots \end{aligned} \tag{5}$$

⋆ indicates equal contribution.

Method	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
LSTM	13.06%	13.20%	14.15%	13.70%	12.85%	12.35%	12.15%	13.00%
WReN	33.97%	58.40%	38.90%	37.70%	21.60%	19.75%	38.85%	22.60%
CNN	36.99%	33.60%	30.30%	33.55%	39.45%	41.25%	43.20%	37.55%
SRAN	45.13%	66.10%	40.70%	38.00%	44.90%	43.20%	47.20%	35.80%
ResNet	53.45%	52.80%	41.85%	44.30%	58.75%	60.15%	63.20%	53.10%
ResNet+DRT	59.59%	58.10%	46.55%	50.40%	65.80%	67.10%	69.10%	60.10%
LEN	71.64%	79.10%	56.05%	60.30%	80.50%	76.40%	79.25%	69.90%
MXGNet	83.99%	94.25%	60.50%	64.85%	96.60%	96.40%	94.05%	81.25%
CoPINet	91.42%	95.05%	77.45%	78.85%	99.10%	99.65%	98.50%	91.35%

Table 1. Testing accuracy of baseline models on RAVEN [4]. All baseline models are trained on the entire dataset. Notations are the same as those in Table 1 in the main text.

Method	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
LSTM	12.22%	11.55%	12.65%	12.80%	12.25%	11.65%	11.95%	12.70%
CNN	12.56%	13.55%	13.35%	12.15%	12.50%	11.80%	11.55%	13.05%
ResNet	18.38%	22.60%	15.45%	18.05%	19.00%	19.55%	17.45%	16.55%
ResNet+DRT	20.74%	24.15%	18.20%	19.75%	21.95%	22.05%	21.00%	18.05%
WReN	21.49%	24.00%	24.95%	20.10%	19.70%	19.85%	21.25%	20.55%
LEN	32.75%	44.80%	27.90%	23.85%	34.05%	34.35%	35.80%	28.50%
MXGNet	33.05%	40.65%	27.85%	24.70%	35.80%	34.45%	36.35%	31.55%
CoPINet	39.09%	44.35%	32.70%	28.50%	42.75%	41.00%	45.20%	39.15%
SRAN	54.41%	66.95%	45.00%	41.30%	60.25%	57.75%	62.70%	46.95%

Table 2. Testing accuracy of baseline models on I-RAVEN [2]. All baseline models are trained on the entire dataset. Notations are the same as those in Table 1 in the main text.

Assuming that based on the distribution of $r^{\texttt{Num}}$, the probabilistic execution engine picks the rule of `Arithmetic plus`, the answer distribution can then be computed as

$$\begin{aligned}
P(\texttt{Num}_9 = 1) &= 0.0 \\
P(\texttt{Num}_9 = 2) &= P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 1) \\
P(\texttt{Num}_9 = 3) &= P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 2) \\
&\quad + P(\texttt{Num}_7 = 2) \times P(\texttt{Num}_8 = 1) \\
P(\texttt{Num}_9 = 4) &= P(\texttt{Num}_7 = 1) \times P(\texttt{Num}_8 = 3) \\
&\quad + P(\texttt{Num}_7 = 2) \times P(\texttt{Num}_8 = 2) \\
&\quad + P(\texttt{Num}_7 = 3) \times P(\texttt{Num}_8 = 1).
\end{aligned} \tag{6}$$

After normalization, we have a generated `Number` distribution

$$P(\texttt{Num}_9) = [0.00, 0.06, 0.40, 0.54]. \tag{7}$$

The generated distribution is compared with the "observed" distribution from each candidate panel using the Jensen–Shannon Divergence (JSD).

2. Rules in RPM

We use the rules summarized in [1]: Constant, Arithmetic, Progression, and Distribute of Three. As rules are parameterized in [2, 4], we have, in total, more than 4 different rule instantiations; see [1, 2, 4] for rule semantics and parameterization.

3. Full Dataset Training

Table 1 shows the testing results of different baseline models trained on the full RAVEN dataset. Table 2 shows the testing results of different baseline models trained on the full I-RAVEN dataset. A full dataset includes all configurations and therefore is 7 times the size of each training regime. As noted in [4] and shown in Tables 4 to 21, incorporating training samples from distinctive configurations will improve a model's performance in general. This observation makes it more impressive that the proposed Probabilistic Abduction and Execution (PrAE) learner, though trained *only on 2x2Grid*, surpasses all models trained *on the entire dataset* during testing on 2x2Grid.

4. Object CNN Architecture

The object CNN in our paper consists of 4 independent branches. All network branches take in the same image region and produce distributions of object attributes of objectiveness, type, size, and color, respectively.

Table 3 shows the LeNet-like architecture [3] we used for each branch. Parameters for convolution denote output

Operator	Parameters
Convolution	$[6, 5, 1]$
BatchNorm	6
SoftPlus	
MaxPool	2
Convolution	$[16, 5, 1]$
BatchNorm	16
SoftPlus	
MaxPool	2
Linear	120
SoftPlus	
Linear	84
SoftPlus	
Linear	m
SoftMax	

Table 3. The architecture used for each branch in the object CNN.

channel size, kernel size, and stride. A batchnorm layer is denoted using its channel size and a max pooling layer its stride. We use the output size to parameterize a fully-connected layer. m is set based on each attribute's dimension: $m = 2$ for objectiveness, $m = 5$ for type, $m = 6$ for size, and $m = 10$ for color.

5. Training on Other Configurations

Tables 4 to 12 show the testing performance of baseline models trained on other configurations in RAVEN, in addition to the 2x2Grid configuration reported in the main text. Tables 13 to 21 show the testing performance of baseline models trained on other configurations in I-RAVEN, in addition to the 2x2Grid configuration. We note that each baseline model's final performance does not vary significantly with respect to different training configurations.

References

[1] Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, 97(3):404, 1990. 2

[2] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[4] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	10.36%	28.80%	10.05%	10.15%	5.70%	4.85%	7.40%	5.60%
3x3Grid	9.77%	9.60%	8.40%	29.10%	5.10%	4.45%	5.30%	6.45%
L-R	8.27%	10.75%	9.20%	9.80%	8.25%	6.70%	7.00%	6.20%
U-D	7.95%	9.20%	9.35%	10.05%	4.95%	9.00%	6.40%	6.70%
O-IC	7.40%	9.35%	8.45%	7.50%	3.85%	4.80%	11.90%	5.95%
O-IG	8.12%	9.45%	8.95%	9.00%	5.95%	4.70%	6.80%	12.00%

Table 4. Testing accuracy of WReN on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	12.80%	12.70%	13.05%	14.35%	12.50%	13.20%	11.40%	12.40%
3x3Grid	12.64%	11.80%	14.35%	13.75%	12.45%	11.90%	12.85%	11.35%
L-R	12.92%	11.10%	13.40%	13.50%	14.45%	14.00%	11.45%	12.55%
U-D	12.41%	12.20%	12.35%	12.60%	11.95%	12.40%	13.30%	12.05%
O-IC	12.15%	10.90%	11.85%	12.60%	12.30%	12.85%	13.10%	11.45%
O-IG	12.46%	12.40%	13.60%	13.05%	11.15%	11.70%	11.90%	13.40%

Table 5. Testing accuracy of LSTM on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	18.10%	60.00%	9.95%	11.00%	7.40%	5.90%	18.90%	13.55%
3x3Grid	12.76%	12.85%	13.65%	39.95%	5.35%	5.45%	6.25%	5.80%
L-R	15.71%	16.10%	11.25%	9.65%	56.80%	4.90%	5.35%	5.90%
U-D	18.09%	21.35%	12.30%	13.65%	5.55%	59.50%	6.40%	7.90%
O-IC	19.86%	18.40%	9.95%	14.00%	4.15%	4.65%	67.75%	20.15%
O-IG	18.56%	17.25%	8.15%	13.05%	5.15%	7.05%	22.20%	57.05%

Table 6. Testing accuracy of LEN on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	14.64%	16.40%	14.65%	13.65%	14.20%	15.45%	14.15%	14.00%
3x3Grid	14.24%	13.60%	15.10%	17.60%	12.30%	13.30%	14.00%	13.80%
L-R	14.56%	15.10%	14.20%	13.75%	17.10%	14.50%	13.55%	13.75%
U-D	15.54%	16.50%	13.90%	15.10%	15.05%	19.00%	15.15%	14.05%
O-IC	16.62%	15.25%	16.50%	13.90%	17.10%	15.60%	19.95%	18.05%
O-IG	15.61%	15.45%	13.45%	15.35%	14.30%	15.05%	17.25%	18.40%

Table 7. Testing accuracy of CNN on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	36.20%	79.60%	26.60%	24.65%	33.25%	37.35%	26.20%	25.75%
3x3Grid	15.03%	9.55%	17.80%	31.10%	11.20%	10.30%	14.10%	11.15%
L-R	36.16%	32.60%	26.25%	29.05%	75.05%	24.95%	35.10%	30.10%
U-D	39.59%	34.05%	26.70%	28.85%	32.40%	82.55%	38.80%	33.75%
O-IC	34.12%	29.60%	22.35%	25.85%	35.05%	30.90%	63.00%	32.10%
O-IG	33.69%	27.35%	26.85%	29.10%	31.45%	29.10%	37.75%	54.25%

Table 8. Testing accuracy of MXGNet on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	24.85%	33.30%	25.30%	22.70%	23.80%	22.95%	23.30%	22.60%
3x3Grid	20.26%	19.30%	20.90%	22.00%	19.75%	19.30%	20.75%	19.85%
L-R	29.66%	26.50%	24.10%	24.05%	40.65%	35.40%	30.55%	26.40%
U-D	30.33%	31.95%	24.40%	26.50%	33.30%	42.35%	28.35%	25.45%
O-IC	27.91%	25.45%	21.80%	23.65%	30.20%	29.10%	34.40%	30.80%
O-IG	24.76%	21.40%	21.75%	22.45%	24.15%	22.30%	30.05%	31.25%

Table 9. Testing accuracy of ResNet on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	33.07%	48.10%	28.55%	26.05%	37.85%	32.45%	35.00%	23.50%
3x3Grid	32.97%	29.05%	35.35%	37.65%	36.35%	36.15%	28.50%	27.75%
L-R	30.53%	27.95%	22.85%	24.10%	40.55%	36.80%	31.10%	30.35%
U-D	33.29%	30.30%	25.90%	29.05%	40.40%	41.25%	34.55%	31.55%
O-IC	31.54%	31.45%	26.00%	27.50%	34.50%	31.20%	39.85%	30.25%
O-IG	28.84%	26.80%	24.15%	25.25%	25.10%	27.50%	34.75%	38.35%

Table 10. Testing accuracy of ResNet+DRT on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	19.71%	60.20%	13.70%	12.85%	11.30%	17.15%	13.90%	8.85%
3x3Grid	18.58%	13.95%	28.55%	54.60%	10.20%	9.55%	6.35%	6.85%
L-R	17.91%	25.35%	14.85%	14.85%	32.75%	16.95%	13.80%	6.85%
U-D	19.26%	27.85%	16.70%	15.40%	20.15%	30.95%	16.15%	7.65%
O-IC	18.61%	22.80%	13.10%	14.10%	15.15%	16.05%	34.95%	14.10%
O-IG	14.07%	14.95%	12.35%	13.25%	7.90%	7.95%	14.60%	27.50%

Table 11. Testing accuracy of SRAN on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	54.77%	88.00%	40.45%	39.10%	63.95%	58.40%	53.50%	40.00%
3x3Grid	50.11%	28.20%	48.10%	66.30%	58.45%	54.85%	53.05%	41.80%
L-R	52.91%	38.35%	42.15%	47.85%	87.40%	68.75%	49.80%	36.10%
U-D	48.74%	42.90%	41.10%	36.05%	68.75%	90.55%	33.70%	28.10%
O-IC	53.75%	27.75%	38.75%	46.20%	65.20%	66.30%	79.50%	52.55%
O-IG	53.79%	36.85%	38.05%	41.60%	60.40%	56.20%	63.10%	80.30%

Table 12. Testing accuracy of CoPINet on different configurations in RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	15.38%	22.10%	13.55%	14.00%	14.00%	14.65%	15.45%	13.90%
3x3Grid	13.86%	14.35%	14.45%	15.35%	13.30%	12.50%	13.25%	13.80%
L-R	13.82%	13.45%	13.64%	14.95%	15.45%	13.15%	13.25%	12.85%
U-D	13.47%	13.80%	13.05%	13.70%	12.45%	16.85%	12.15%	12.30%
O-IC	14.43%	14.40%	12.85%	14.35%	12.25%	13.15%	17.70%	16.30%
O-IG	14.60%	13.75%	12.65%	14.30%	13.40%	14.25%	16.35%	17.50%

Table 13. Testing accuracy of WReN on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	12.87%	12.60%	13.55%	11.75%	12.20%	13.85%	13.45%	12.70%
3x3Grid	12.53%	12.20%	11.15%	12.20%	12.25%	13.05%	14.00%	12.85%
L-R	12.34%	11.95%	12.40%	13.40%	11.70%	12.60%	11.70%	12.60%
U-D	11.85%	12.35%	11.60%	11.85%	12.65%	10.65%	12.30%	11.55%
O-IC	12.59%	12.95%	13.10%	12.45%	11.70%	12.45%	12.15%	13.35%
O-IG	12.46%	13.40%	11.65%	12.00%	11.35%	12.55%	12.70%	13.55%

Table 14. Testing accuracy of LSTM on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	15.28%	25.05%	14.15%	13.95%	13.15%	13.75%	13.80%	13.10%
3x3Grid	13.23%	13.55%	13.50%	15.65%	12.45%	12.25%	12.85%	12.35%
L-R	13.86%	16.10%	13.50%	13.20%	16.70%	12.10%	12.65%	12.80%
U-D	14.60%	16.60%	14.45%	14.40%	14.00%	17.25%	12.70%	12.80%
O-IC	14.31%	14.35%	14.30%	13.20%	12.55%	12.35%	17.65%	15.75%
O-IG	14.28%	14.10%	15.05%	13.70%	12.90%	13.65%	14.05%	16.50%

Table 15. Testing accuracy of LEN on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	12.21%	14.20%	12.40%	10.95%	10.90%	11.55%	12.00%	13.50%
3x3Grid	12.53%	12.25%	13.40%	12.25%	12.55%	13.40%	12.20%	11.65%
L-R	12.46%	12.90%	12.55%	12.30%	11.85%	11.70%	13.80%	12.15%
U-D	12.51%	12.00%	13.70%	12.25%	12.15%	10.85%	13.65%	12.95%
O-IC	12.94%	13.75%	12.75%	12.55%	13.00%	12.80%	12.70%	13.05%
O-IG	12.41%	11.95%	12.25%	12.45%	12.15%	11.45%	12.55%	14.05%

Table 16. Testing accuracy of CNN on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	15.54%	20.70%	14.80%	15.05%	17.00%	17.00%	15.85%	15.35%
3x3Grid	14.48%	14.95%	15.70%	15.70%	13.30%	14.65%	12.70%	14.35%
L-R	12.15%	12.00%	11.75%	13.35%	11.30%	13.85%	12.30%	10.50%
U-D	12.88%	13.20%	12.60%	12.20%	12.25%	13.55%	13.00%	13.35%
O-IC	12.92%	12.20%	12.45%	13.55%	12.90%	13.35%	14.30%	11.70%
O-IG	12.34%	12.70%	12.50%	13.05%	12.60%	12.65%	11.70%	11.15%

Table 17. Testing accuracy of MXGNet on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	14.75%	19.40%	14.70%	13.55%	14.80%	13.45%	14.55%	12.80%
3x3Grid	12.86%	13.00%	13.50%	13.50%	12.75%	13.00%	11.30%	13.00%
L-R	13.67%	13.80%	13.80%	14.60%	14.05%	12.25%	13.45%	13.75%
U-D	13.39%	12.95%	13.95%	13.40%	13.35%	13.05%	14.10%	12.90%
O-IC	13.13%	13.00%	13.90%	12.90%	12.75%	12.90%	13.45%	13.00%
O-IG	13.03%	13.45%	13.25%	12.40%	13.65%	12.00%	13.75%	12.70%

Table 18. Testing accuracy of ResNet on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	14.94%	19.75%	15.30%	14.55%	13.50%	13.85%	14.90%	12.75%
3x3Grid	13.06%	13.85%	12.65%	13.70%	11.80%	12.85%	13.15%	13.40%
L-R	13.33%	14.05%	13.45%	13.15%	13.55%	13.55%	12.45%	13.10%
U-D	12.74%	13.80%	12.85%	12.00%	13.15%	13.45%	12.10%	11.85%
O-IC	12.48%	12.75%	13.15%	12.65%	12.35%	11.60%	13.25%	11.60%
O-IG	13.04%	12.75%	13.20%	14.80%	12.45%	12.40%	13.30%	12.40%

Table 19. Testing accuracy of ResNet+DRT on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	32.01%	63.65%	20.60%	20.60%	30.15%	35.80%	30.75%	22.50%
3x3Grid	26.59%	31.10%	30.60%	31.50%	25.20%	22.60%	24.55%	20.55%
L-R	30.70%	39.60%	22.25%	19.85%	48.25%	32.20%	32.20%	20.55%
U-D	31.25%	37.95%	21.40%	18.10%	37.10%	49.80%	32.20%	22.20%
O-IC	31.26%	34.15%	20.45%	20.70%	30.20%	32.90%	50.95%	29.50%
O-IG	24.54%	21.60%	19.10%	19.45%	20.10%	18.80%	30.60%	42.10%

Table 20. Testing accuracy of SRAN on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.

Training Regime	Acc	Center	2x2Grid	3x3Grid	L-R	U-D	O-IC	O-IG
Center	23.79%	39.25%	23.65%	20.40%	21.45%	22.15%	19.30%	20.35%
3x3Grid	22.64%	22.95%	25.25%	26.55%	21.30%	21.30%	21.45%	19.65%
L-R	24.65%	26.15%	21.20%	20.90%	33.30%	28.50%	21.45%	21.05%
U-D	24.38%	26.75%	21.95%	22.10%	26.20%	33.85%	21.85%	17.95%
O-IC	23.15%	22.25%	17.10%	20.65%	20.65%	22.55%	34.65%	26.70%
O-IG	23.21%	19.05%	20.75%	23.15%	17.55%	19.75%	27.50%	34.70%

Table 21. Testing accuracy of CoPINet on different configurations in I-RAVEN. The model is trained on configurations listed in rows and tested on all configurations. Notations are the same as those in Table 1 in the main text.