
SymNet 2.0: Effectively handling Non-Fluents and Actions in Generalized Neural Policies for RDDDL Relational MDPs Supplementary material

Vishal Sharma¹

Daman Arora¹

Florian Geißer²

Mausam¹

Parag Singla¹

¹Indian Institute of Technology Delhi {vishal.sharma, cs5180404, mausam, parags}@cse.iitd.ac.in

²Independent Researcher, {florian.geisser.work}@gmail.com

A APPENDIX

A.1 PROOFS OF PROPOSITIONS

Proposition 1. *Let there be two actions $a(o_1)$ and $a(o_2)$ of same action type $\text{type}(a)$ where $o_1 \neq o_2$. Let both these actions effect same set of state-variables i.e. $\mathcal{P}_{a(o_1)} = \mathcal{P}_{a(o_2)}$. Then, the scores computed by SymNet for both of these actions will be same.*

Proof. The proof comes from the fact that, SymNet’s action decoder considers the node embeddings of only those nodes that the action effects (and global embedding) while ignores the actual parameters of the action. Hence, $\text{score}(a(o_1)) = \text{AD}_{\text{type}(a)}(\text{maxpool}_{P \in \mathcal{P}_{a(o_1)}}(\text{ne}(\text{args}(P))), \text{ge}) = \text{score}(a(o_2))$ as $\mathcal{P}_{a(o_1)} = \mathcal{P}_{a(o_2)}$.

Proposition 2. *Let u and v be two nodes with label o_u and o_v corresponding to object tuples of some state-variables in \mathcal{G}_{sym} . Let $d_{\text{sym}}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in \mathcal{G}_{sym} and let $d_{\text{sym}2}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in $\mathcal{G}_{\text{sym}2}$. Then, $d_{\text{sym}2}(u, v) \leq d_{\text{sym}}(u, v)$.*

Proof. If there exists a path of length l between u and v in \mathcal{G}_{sym} then there will be a path between u and v in $\mathcal{G}_{\text{sym}2}$ because by construction $\mathcal{G}_{\text{sym}2}$ contains all nodes and edges of \mathcal{G}_{sym} . Hence, $d_{\text{sym}2}(u, v)$ is at least equal to $d_{\text{sym}}(u, v)$. Next, if the labels o_u and o_v share any object \tilde{o} as parameter then there will bi-directional edges $e(u, \tilde{o})$ and $e(\tilde{o}, v)$ in the position based graphs of \mathcal{G}_{sym} . Hence, the distance $d_{\text{sym}2}(u, v) = \min(l, 2) \leq d_{\text{sym}}(u, v)$.

A.2 ARCHITECTURE DETAILS

SymNet2.0 construct a multi-graph with nodes created on objects and tuples, and edges based on connections in DBNs, affects of actions, and object position in tuples as described in the main paper. For

each node v , node embedding is computed using Graph Attention Networks Veličković et al. [2018] as $\bar{v} = \text{mlp}(\text{concat}(GAT_1(G_1), \dots, GAT_{|N|}(G_{|N|})))$ where N is the number of adjacencies. In our experiments in the main paper, we use a GAT of depth 1, with 8 attention heads. For all domains, the dimension of the final node embedding of all nodes is 20. To capture the global view of the state, a state embedding is computed as $\bar{s} = \text{maxpool}(\bar{v})$. For each action template, we use an action decoder which is an MLP with 1 hidden layer of dimension 20. Finally, scores of all action scores are normalized using softmax to get a policy.

Our model is trained using Adam with a learning rate of 3×10^{-3} on a batch size of 40 for 500 epochs with validation being done every 50 epochs. We validate on all the checkpoints by performing 200 rollouts on the validation instances.

All experiments are done on a system with Intel Xeon CPU E5-2680 v3(2.50 GHz) processor with 62 GB RAM and a NVIDIA Tesla K40 GPU.

A.3 DOMAIN DESCRIPTION

- Triangle Tireworld (TT):** Triangle tireworld consists of a triangle shaped maze out of which a fixed set of cells are equipped with a spare tire. An agent must navigate to the goal. Each transition could result in a flat-tire. In case the agent doesn’t have a spare-tire, it can’t navigate further. Spare-tires can be picked up from cells which have a spare tire.
- Crossing Traffic (CT):** Crossing traffic requires a robot to navigate in a gridworld from a start position(S) in the south-east to a goal position(G) north-east. There is a constant flow of traffic from the east to west. Landing on the same cell as a traffic object results in death.
- Academic Advising (Acad):** In the Academic advising domain, a student is to complete a set of requirement courses from a set of available courses. A course might have 0 or more prerequisites and the probability of

IPPC Test Instances 5-10

Model	GAT	TT	CT	Acad	Elev	Tam	Nav	GoL	Skill	Sys	Wild	Traffic	Recon	Mean
PROST	-	0.52	0.86	0.47	1.00	0.94	0.88	1.00	1.00	0.64	0.66	1.00	0.99	0.83
SymNet	1	0.00	0.37	0.58	0.31	0.54	0.53	0.20	-0.40	0.61	0.26	0.00	0.03	0.25
SymNet-IL	1	0.68	0.91	0.71	0.38	0.62	0.56	0.20	-0.50	0.49	0.68	-0.18	0.03	0.38
SymNet-IL	2	0.47	0.68	0.76	0.41	0.58	0.62	0.07	-0.37	0.70	0.75	-0.34	0.03	0.36
SymNet-IL	3	0.50	0.38	0.77	0.38	0.59	0.41	0.19	-0.50	0.84	0.92	-0.37	0.03	0.35
SYMNET2.0	1	0.70	0.95	0.81	0.44	0.91	0.47	0.31	0.43	0.92	0.73	0.28	0.30	0.60
SYMNET2.0	2	0.67	0.81	0.83	0.58	0.88	0.47	0.24	0.47	0.91	0.73	-0.23	0.46	0.57
SYMNET2.0	3	0.70	0.57	0.78	0.36	0.85	0.47	0.18	0.13	0.86	0.53	0.02	0.31	0.48

Table 2: Results showing comparison between SYMNET2.0 and the baselines on 12 IPPC domains when we vary the neighborhood size of GAT (column "GAT").

completing a course increases with the number of completed prerequisite courses. A penalty is provided if a course is repeated. Therefore, one must complete all courses as soon as possible to attain the maximum reward.

4. **Elevators (Elev):** In the Elevators domain, there are multiple elevators in a building with multiple floors. Passengers can arrive on different floors with different probabilities. The agent must ensure that passengers wait for the least amount of time on their floors.
5. **Tamarisk (Tam):** In Tamarisk, tamarisk(a shrub) spreads downstream and upstream(with lower probability). The shrub must be eradicated, or native species must be planted at those locations. The cost of eradication of tamarisk, and restoration of the native species must be minimized by the agent.
6. **Navigation (Nav):** In the navigation domain, a robot must navigate in a rectangular grid-world to the goal position. Each cell has a predefined probability of death. The robot must increase its chances of survival by navigating through a low-risk path as well as minimize the path length in order to attain a high reward.
7. **Game of Life (GoL):** The Game of life domain consists of a grid with a specific set of cells alive at a particular time step. In a move, the agent can SET a particular grid cell. If a cell has 0 or 1 alive neighbours, then it dies with high probability. If there are 2 or 3 neighbours, then it lives with high probability. If there are more than 3 neighbours, then it dies with high probability.
8. **Skill Teaching (Skill):** The skill teaching domain contains various skills with varying weights. A skill can have 0 or more prerequisite skills. There are two levels of proficiencies for each skill: medium and high. Giving a hint for a skill increases its proficiency to medium assuming the student has high proficiency in all prerequisite skills. The agent can also ask a question for a particular skill. If the student answers correctly, then

the proficiency becomes high in the particular skill. Proficiency can decrease stochastically as well. The reward is proportional to the weight of the skills with high/medium-proficiency.

9. **Sysadmin (Sys):** In Sysadmin, each instance has a set of various computers connected in a fixed topology for each instance. The reward obtained per move increases with the number of computers which are ON at a particular time step. An OFF computer can turn ON with a small probability and an ON computer can turn OFF with a probability which increases with the number of neighbours which are in the ON state.
10. **Wildfire (Wild):** In the wildfire domain, fires are spreading through an entire grid and one needs to minimize the number of cells on fire. The probability of fire spreading increases with the number of neighbours which are on fire. Actions involve either putting out the fire or removing the fuel at any particular cell.
11. **Traffic:** The traffic domain involves 2 horizontal and 2 vertical roads. On all 4 intersections, there is a traffic light which needs to be controlled. The goal of the agent is to minimize congestion(two cars in two consecutive cells). The inflow of traffic is only from one of the ends of the 4 roads, and the flow of the traffic is specified in the instance file.
12. **Recon:** Recon involves an agent which is equipped with tools to capture pictures, detect life, and detect water. To obtain reward, the agent must take pictures where life was detected. On reporting negative results, the tools can contaminate the object they were used on. Therefore, the agent must understand which tools are to be used, whether they must move through hazards in the grid, and whether they should be repaired.

A.4 STATISTICAL SIGNIFICANCE TEST

We ran the paired T-test¹ to examine the statistical significance of gain in rewards obtained by SYMNET2.0 vs SymNet-IL (better performing SymNET variation). For each of the 6 test instances of each of the 12 domains, we compute the mean reward over 3 runs for SymNet and SYMNET2.0, resulting in 72 paired samples. Next, using the paired T-test we reject the null hypothesis that the mean of the distribution (over 72 points) for SymNet-IL is greater than the mean of the distribution (over 72 points) for SYMNET2.0 with p-value of 0.9994.

A.5 LARGE INSTANCE GENERATION

We generate 4 instances of increasing sizes for all 12 domains. For this, we use generators provided by official Repository of RDDDL Simulator by Scott Sanner². The exact generation script used for generation will be shared in the final version of the paper. Some of the important parameters of each domain are given below. We keep all other parameters as close as possible to values seen on instances 1-10 for that particular domain. Table 3 shows the sizes of each instance in terms of number of state-variables.

1. **Acad**: Number of courses are 36, 48, 70, 90
2. **CT**: Width of grids are 7, 8, 8, 3.
Corresponding heights are 8, 9, 11, 20
3. **Nav**: Width of grids are 15, 12, 20, 30.
Corresponding heights are 9, 8, 5, 3
4. **Sys**: Number of computers are 60, 75, 100, 120
5. **GOL**: Width of grids are 6, 7, 8, 9.
Corresponding heights are 8, 9, 10, 12.
6. **Wild**: Width of grids are 11, 12, 13, 15.
Corresponding heights are 5, 4, 5, 4.
7. **Skill**: Number of skills are 12, 14, 16, 18
8. **Traffic**: Number of cells are 84, 84, 98, 108
9. **Tamarisk**: Number of reaches are 9, 9, 10, 10. Corresponding number of slots are 2, 3, 2, 3
10. **Elev**: Number of elevators are 2, 1, 2, 3.
Corresponding number of floors are 6, 8, 8, 10
11. **Recon**: Width and height of grids are 6, 8, 10, 9.
Corresponding number of objects are 7, 8, 8, 7
12. **TT**: Grid sizes are 91, 120, 153, 190

A.6 STANDARD ERROR AMONG RESULTS

Tables 4 and 6 show the mean relative score and standard error across 3 runs of each model. The standard error is somewhat high in some cases as only 3 runs were done due to resource constraints. Tables 5 and 7 show scores achieved on individual runs for IPPC and large instances.

A.7 RAW REWARDS

Tables 8 to 19 show the raw long term rewards for all 12 domains.

References

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018.

¹https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

²<https://github.com/ssanner/rddlsim/tree/master/src/rddl/competition/generators>

	IPPC Train			IPPC Test							Large			
Domain	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Acad	20	20	30	30	40	40	50	50	60	60	72	96	140	180
CT	12	12	24	24	40	40	60	60	84	84	98	128	160	114
Elev	9	16	16	12	20	20	15	24	24	18	28	24	36	58
GoL	9	9	9	16	16	16	25	25	25	30	64	81	100	144
Nav	12	15	20	30	30	40	50	60	80	100	135	96	100	90
Sys	10	10	20	20	30	30	40	40	50	50	60	75	100	120
Wild	18	18	32	32	50	50	60	60	72	72	110	96	130	120
Traffic	32	32	44	44	56	56	68	68	80	80	92	92	104	116
Tam	16	24	20	30	24	36	28	42	32	48	36	54	40	60
TT	12	12	27	27	48	48	75	75	108	108	147	192	243	300
Skill	12	12	24	24	36	36	42	42	48	48	66	84	84	96
Recon	30	30	41	41	54	54	54	69	69	69	80	114	150	125

Table 3: Number of state-variables (SP_O) per instance for IPPC and large domains

IPPC Test Instances 5-10						
Model	TT	CT	Acad	Elev	Tam	Nav
r_1 PROST	0.53	0.86	0.47	1.00	0.94	0.88
r_2 SymNet	0.00±0.00	0.37±0.01	0.58±0.03	0.31±0.00	0.55±0.01	0.53±0.03
r_3 SymNet-IL	0.83±0.05	0.91±0.02	0.72±0.13	0.38±0.08	0.63±0.02	0.56±0.01
r_4 SYMNET2.0	0.81±0.06	0.95±0.02	0.82±0.04	0.44±0.09	0.92±0.01	0.47±0.00
Model	GoL	Skill	Sys	Wild	Traffic	Recon
r_1 PROST	1.00	1.00	0.65	0.70	1.00	0.99
r_2 SymNet	0.20±0.02	-0.40±0.12	0.62±0.03	0.27±0.13	0.00±0.00	0.03±0.00
r_3 SymNet-IL	0.20±0.02	-0.50±0.00	0.49±0.04	0.72±0.19	-0.18±0.17	0.03±0.00
r_4 SYMNET2.0	0.29±0.01	0.43±0.13	0.94±0.01	0.77±0.09	0.28±0.06	0.30±0.06

Table 4: Results showing comparison between SYMNET2.0 and the baselines on 12 IPPC domains. All models are trained on (smaller) instances 1-3 and validated on instance 4. All Rows show results on IPPC instances 5-10. Bold values show the best performer among all neural models. Each entry gives the mean relative score \pm standard error over 3 runs.

IPPC Test Instances 5-10						
Model	TT	CT	Acad	Elev	Tam	Nav
r_1 PROST	0.53	0.86	0.47	1.00	0.94	0.88
r_2 SymNet	0.00/0.00/0.00	0.39/0.35/0.37	0.65/0.51/0.58	0.31/0.31/0.30	0.56/0.56/0.52	0.53/0.59/0.48
r_3 SymNet-IL	0.70/0.89/0.91	0.89/0.89/0.95	0.81/0.40/0.95	0.27/0.59/0.29	0.58/0.65/0.66	0.56/0.58/0.54
r_4 SYMNET2.0	0.95/0.71/0.78	0.97/0.98/0.89	0.84/0.72/0.89	0.28/0.65/0.38	0.92/0.94/0.90	0.47/0.48/0.47
Model	GoL	Skill	Sys	Wild	Traffic	Recon
r_1 PROST	1.00	1.00	0.65	0.70	1.00	0.99
r_2 SymNet	0.25/0.19/0.16	-0.10/-0.59/-0.50	0.69/0.58/0.59	0.58/0.04/0.19	0.00/-0.01/0.01	0.03/0.03/0.03
r_3 SymNet-IL	0.20/0.17/0.24	-0.50/-0.50/-0.50	0.57/0.40/0.51	0.96/0.93/0.26	-0.39/-0.39/0.23	0.03/0.03/0.03
r_4 SYMNET2.0	0.28/0.29/0.31	0.12/0.57/0.61	0.97/0.91/0.95	0.64/0.68/0.98	0.26/0.16/0.41	0.17/0.40/0.33

Table 5: Results showing scores of 3 individual runs of SYMNET2.0 and other baselines on 12 IPPC domains on instances 5-10. All models are trained on (smaller) instances 1-3 and validated on instance 4. Each entry shows the relative score on three runs.

Larger Instances						
Model	TT	CT	Acad	Elev	Tam	Nav
r_1 PROST	0.09	0.55	0.39	1.00	0.90	0.44
r_2 SymNet	0.00 \pm 0.00	0.14 \pm 0.01	0.60 \pm 0.05	0.15 \pm 0.02	0.43 \pm 0.02	0.41 \pm 0.19
r_3 SymNet-IL	0.96\pm0.02	0.62 \pm 0.05	0.63 \pm 0.10	0.22\pm0.05	0.52 \pm 0.01	0.19 \pm 0.01
r_4 SYMNET2.0	0.95 \pm 0.03	0.89\pm0.08	0.77\pm0.07	0.19 \pm 0.03	0.94\pm0.03	0.95\pm0.02
Model	GoL	Skill	Sys	Wild	Traffic	Recon
r_1 PROST	0.91	1.00	0.36	1.00	1.00	0.78
r_2 SymNet	0.60 \pm 0.01	-0.82 \pm 0.02	0.51\pm0.10	0.09 \pm 0.03	0.25 \pm 0.00	0.02 \pm 0.00
r_3 SymNet-IL	0.25 \pm 0.17	-0.79 \pm 0.00	-0.65 \pm 0.08	0.22\pm0.09	0.03 \pm 0.10	0.02 \pm 0.00
r_4 SYMNET2.0	0.84\pm0.03	0.34\pm0.13	0.46 \pm 0.35	0.20 \pm 0.04	0.39\pm0.08	0.32\pm0.08

Table 6: Results showing comparison between SYMNET2.0 and the baselines on 12 IPPC domains. All models are trained on (smaller) instances 1-3 and validated on instance 4. All Rows show results on larger instances (11-14) than those in the IPPC. Bold values show the best performer among all neural models. Each entry gives the mean relative score \pm standard error over 3 runs. The standard error is high (only) in some cases as only 3 runs were done due to resource constraints.

Larger Instances						
Model	TT	CT	Acad	Elev	Tam	Nav
r_1 PROST	0.09	0.55	0.39	1.00	0.90	0.44
r_2 SymNet	0.00/0.00/0.00	0.14/0.13/0.16	0.67/0.49/0.65	0.13/0.14/0.19	0.39/0.45/0.44	0.87/0.18/0.17
r_3 SymNet-IL	0.97/1.00/0.92	0.57/0.55/0.74	0.61/0.43/0.85	0.13/0.33/0.21	0.49/0.55/0.53	0.17/0.21/0.19
r_4 SYMNET2.0	1.00/0.89/0.97	0.99/0.99/0.68	0.77/0.62/0.92	0.13/0.22/0.23	0.99/0.96/0.88	0.90/0.97/0.98
Model	GoL	Skill	Sys	Wild	Traffic	Recon
r_1 PROST	0.91	1.00	0.36	1.00	1.00	0.78
r_2 SymNet	0.61/0.57/0.62	-0.79/-0.87/-0.79	0.76/0.33/0.45	0.17/0.03/0.07	0.25/0.26/0.24	0.02/0.02/0.02
r_3 SymNet-IL	-0.14/0.59/0.30	-0.79/-0.79/-0.79	-0.45/-0.74/-0.75	0.35/0.31/0.00	-0.10/-0.10/0.28	0.02/0.02/0.02
r_4 SYMNET2.0	0.90/0.81/0.80	0.01/0.48/0.52	0.98/0.77/-0.38	0.14/0.15/0.31	0.40/0.22/0.55	0.17/0.28/0.52

Table 7: Results showing scores of 3 individual runs of SYMNET2.0 and other baselines on larger instances (11-14) of 12 IPPC domains. All models are trained on (smaller) instances 1-3 and validated on instance 4. Each entry shows the relative score on three runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-41.96	-55.09/-200.0/-81.38	-87.61/-40.67/-61.9	-40.71/-63.01/-41.23
2	-72.8	-94.14/-200.0/-135.46	-75.22/-82.78/-77.77	-75.92/-76.16/-75.48
3	-42.21	-56.56/-200.0/-48.61	-40.27/-39.23/-39.85	-39.51/-42.05/-40.41
4	-200.74	-125.38/-200.0/-133.56	-123.06/-127.14/-121.05	-117.83/-114.4/-115.62
5	-203.19	-148.46/-200.0/-181.06	-143.46/-199.47/-123.53	-150.51/-125.45/-126.68
6	-203.1	-147.78/-200.0/-190.21	-110.5/-124.92/-110.68	-107.04/-115.51/-107.76
7	-201.25	-144.62/-200.0/-139.31	-141.94/-231.42/-121.61	-112.61/-125.04/-113.56
8	-215.43	-191.99/-200.0/-195.19	-160.27/-167.09/-157.07	-161.08/-228.88/-176.38
9	-201.74	-198.23/-200.0/-201.74	-199.74/-256.45/-179.72	-196.79/-237.17/-164.02
10	-202.89	-230.15/-200.0/-220.75	-203.62/-253.51/-179.53	-204.0/-175.65/-211.23
11	-503.44	-214.22/-500.0/-232.3	-247.6/-205.36/-219.74	-167.58/-182.37/-160.47
12	-544.31	-252.3/-500.0/-286.7	-202.83/-282.67/-217.67	-542.3/-223.0/-215.58
13	-528.48	-565.31/-500.0/-545.46	-566.07/-627.13/-453.65	-484.64/-628.01/-466.75
14	-530.59	-507.89/-500.0/-532.79	-588.44/-669.26/-492.84	-377.23/-507.3/-431.3

Table 8: **Acad**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-4.37	-14.54/-12.45/-13.78	-4.49/-4.46/-4.39	-4.29/-4.525/-4.61
2	-5.48	-25.56/-22.71/-25.18	-5.54/-5.32/-5.56	-5.29/-5.62/-5.79
3	-6.08	-24.46/-22.05/-21.32	-5.96/-5.95/-5.97	-5.96/-6.22/-6.08
4	-10.41	-35.38/-34.45/-33.52	-9.74/-10.36/-9.71	-10.68/-9.705/-11.89
5	-7.08	-22.18/-22.0/-21.46	-7.85/-7.97/-7.7	-7.33/-7.15/-8.16
6	-10.03	-32.26/-31.9/-31.9	-15.44/-15.13/-12.38	-10.83/-12.745/-13.01
7	-9.8	-25.12/-29.15/-24.77	-11.16/-10.79/-8.34	-8.79/-8.585/-12.36
8	-25.53	-35.1/-37.38/-35.98	-20.41/-21.91/-17.02	-16.91/-15.545/-17.09
9	-8.43	-18.58/-20.45/-23.0	-9.96/-9.21/-9.35	-8.78/-8.665/-13.96
10	-24.67	-34.22/-34.39/-34.05	-16.87/-17.07/-16.44	-14.95/-13.92/-17.72
11	-44.51	-73.03/-82.33/-65.59	-23.89/-20.21/-15.93	-13.29/-11.14/-33.73
12	-53.95	-96.32/-90.8/-93.56	-51.83/-60.58/-29.09	-26.83/-27.02/-48.23
13	-77.19	-83.8/-87.4/-89.2	-46.67/-36.68/-25.79	-18.99/-21.04/-41.77
14	-56.59	-99.19/-97.57/-99.19	-93.12/-100.0/-92.42	-36.55/-36.29/-64.91

Table 9: **CT**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-45.09	-66.7/-65.97/-66.59	-66.31/-46.63/-43.54	-65.94/-43.02/-45.25
2	-21.62	-55.14/-55.44/-54.99	-55.16/-29.95/-20.07	-55.12/-24.25/-26.66
3	-63.36	-71.36/-71.0/-71.53	-71.04/-63.95/-58.66	-71.28/-64.66/-62.12
4	-55.26	-98.8/-96.25/-99.81	-100.2/-72.45/-88.45	-100.72/-70.48/-76.8
5	-67.6	-110.24/-108.34/-109.94	-110.12/-85.91/-106.83	-112.83/-82.75/-100.37
6	-84.03	-121.63/-121.35/-122.65	-120.31/-100.44/-121.94	-122.3/-100.11/-120.99
7	-80.53	-133.36/-131.94/-132.9	-134.91/-108.89/-126.01	-131.66/-106.48/-118.32
8	-88.83	-145.31/-144.28/-144.81	-147.57/-122.79/-149.62	-146.69/-119.63/-144.32
9	-109.07	-161.69/-159.74/-160.24	-163.85/-139.32/-172.69	-162.69/-137.22/-167.34
10	-66.92	-116.39/-120.07/-119.05	-125.98/-110.45/-119.38	-122.0/-101.5/-107.93
11	-51.88	-280.11/-273.91/-261.65	-282.15/-209.75/-244.61	-277.08/-231.28/-223.89
12	-93.61	-348.93/-341.3/-327.0	-346.32/-298.22/-324.89	-358.4/-319.87/-338.5
13	-43.98	-232.88/-221.14/-228.49	-235.0/-192.83/-214.75	-226.79/-220.82/-212.71
14	-84.29	-262.84/-269.9/-244.59	-258.67/-222.21/-257.54	-258.8/-260.46/-249.56

Table 10: **Eleve**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	202.18	91.55/106.15/112.52	105.52/116.42/154.41	137.84/145.71/143.5
2	127.66	88.66/75.08/77.56	81.94/89.53/104.66	88.68/102.405/115.8
3	150.05	110.14/103.61/103.69	115.14/118.33/130.16	116.14/123.03/128.96
4	363.98	242.13/220.33/208.28	220.28/202.5/212.37	218.12/217.865/226.89
5	317.5	234.75/222.09/219.0	230.06/222.03/234.27	231.31/246.42/237.94
6	280.25	247.56/241.52/237.97	244.88/240.78/247.74	244.91/251.935/249.33
7	520.26	288.64/311.08/290.29	295.65/311.37/307.01	346.06/354.07/340.7
8	463.43	337.13/328.37/325.36	310.04/311.39/327.17	336.5/347/336.56
9	432.52	352.95/338.93/341.29	326.22/340.35/337.19	349.51/355.825/349.88
10	616.26	233.31/206.06/197.42	305.0/194.53/270.5	272.82/210.73/278.63
11	1879.75	1716.52/1691.64/1701.29	1545.72/1666.89/1645.75	1779.7/1843.89/1759.15
12	2303.29	2219.74/2220.68/2203.26	2057.16/2243.91/2139.99	2283.78/2272.0/2243.02
13	2883.13	2850.49/2819.26/2829.1	2690.04/2831.26/2763.24	2919.32/2901.4/2902.16
14	4057.8	4027.7/4040.2/4081.85	3825.65/4041.57/3978.15	4097.75/4009.19/4084.57

Table 11: **GoL**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-9.28	-9.44/-9.12/-10.56	-10.56/-9.76/-10.4	-8.48/-9.44/-9.6
2	-11.2	-11.35/-10.6/-11.35	-10.45/-10.75/-11.5	-10.9/-10.9/-11.05
3	-14.48	-14.19/-13.76/-13.18	-12.45/-13.61/-13.18	-14.48/-13.03/-13.76
4	-17.33	-16.91/-17.05/-16.24	-16.91/-16.78/-16.24	-16.38/-17.73/-16.51
5	-20.21	-30.65/-30.31/-20.1	-33.37/-29.8/-31.84	-20.6/-20.2/-20.6
6	-22.9	-37.69/-37.2/-22.43	-37.03/-36.87/-37.69	-21.66/-21.86/-21.86
7	-24.62	-39.52/-39.2/-22.54	-39.52/-39.36/-39.52	-23.35/-22.72/-22.9
8	-30.45	-32.09/-31.32/-40.0	-31.11/-30.9/-30.9	-40.0/-40.0/-40.0
9	-35.73	-34.67/-34.41/-40.0	-34.6/-34.54/-34.8	-40.0/-40.0/-40.0
10	-37.78	-37.36/-36.94/-40.0	-36.82/-37.36/-37.48	-40.0/-40.0/-40.0
11	-88.13	-55.2/-100.0/-100.0	-100.0/-100.0/-100.0	-54.56/-51.36/-48.16
12	-64.05	-43.91/-100.0/-100.0	-100.0/-100.0/-100.0	-43.91/-38.23/-43.2
13	-78.11	-51.86/-100.0/-100.0	-100.0/-100.0/-99.06	-53.02/-48.96/-46.06
14	-79.01	-68.08/-71.88/-73.4	-73.4/-66.56/-71.12	-62.8/-61.2/-61.6

Table 12: **Nav**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	3.32	0.0/0.0/0.0	0.0/0.0/0.0	1.72/1.96/2.41
2	3.07	0.0/0.0/0.0	0.0/0.0/0.0	2.43/2.44/2.33
3	14.02	0.0/0.0/0.0	0.0/0.0/0.0	12.74/12.56/12.46
4	2.68	0.0/0.0/0.0	0.0/0.0/0.0	2.29/1.93/1.33
5	14.41	0.0/0.0/0.0	0.0/0.0/0.0	1.18/8.11/11.84
6	10.39	0.0/0.0/0.0	0.0/0.0/0.0	0.86/0.0/0.79
7	5.04	0.0/0.0/0.0	0.0/0.0/0.0	4.04/0.0/5.29
8	10.12	0.0/0.0/0.0	0.0/0.0/0.0	0.0/7.28/0.0
9	9.53	0.0/0.0/0.0	0.0/0.0/0.0	0.0/9.49/0.0
10	5.65	0.0/0.0/0.0	0.0/0.0/0.0	0.0/0.0/0.0
11	3.46	0.0/0.0/0.0	0.0/0.0/0.0	0.0/9.19/8.85
12	12.49	0.0/0.0/0.0	0.0/0.0/0.0	0.0/0.0/1.06
13	17.58	0.0/0.0/0.0	0.0/0.0/0.0	0.0/0.0/0.0
14	21.99	0.0/0.0/0.0	0.0/0.0/0.0	19.54/2.79/30.78

Table 13: **Recon**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	66.95	59.17/65.78/60.77	64.64/67.35/65.55	67.67/66.94/65.93
2	78.77	71.12/78.26/73.17	77.21/80.06/78.77	77.44/77.75/78.16
3	91.03	37.7/53.89/62.55	-219.21/-219.21/-219.21	89.4/93.1/91.11
4	102.9	65.24/57.39/52.59	-231.67/-231.67/-231.67	92.58/91.87/100.44
5	10.21	-66.04/-455.87/-406.81	-406.81/-406.81/-406.81	-333.32/-189.85/-4.13
6	6.07	-102.07/-490.74/-490.74	-490.74/-490.74/-490.74	-64.4/-64.02/-154.21
7	-65.82	-609.23/-652.96/-609.23	-609.23/-609.23/-609.23	-498.65/-209.87/-251.2
8	-182.39	-724.82/-770.66/-724.82	-724.82/-724.82/-724.82	-525.78/-444.15/-416.87
9	-152.77	-695.75/-695.75/-695.75	-695.75/-695.75/-695.75	-452.55/-259.99/-222.86
10	-238.24	-849.75/-899.57/-849.75	-849.75/-849.75/-849.75	-596.85/-342.5/-414.96
11	-604.37	-2855.3/-2995.11/-2855.3	-2855.3/-2855.3/-2855.3	-2002.95/-675.87/-1337.36
12	-957.54	-4043.19/-4160.32/-4043.19	-4043.19/-4043.19/-4043.19	-2436.21/-2508.69/-2112.47
13	-1269.85	-5258.32/-5361.75/-5258.32	-5258.32/-5258.32/-5258.32	-3826.92/-2621.45/-1806.55
14	-1987.77	-5292.95/-5402.39/-5292.95	-5292.95/-5292.95/-5292.95	-3517.04/-3009.92/-2759.91

Table 14: **Skill**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	339.87	331.88/325.05/323.23	328.1/331.48/327.17	328.47/332.27/331.69
2	307.46	292.02/283.39/278.47	292.38/289.83/277.01	292.57/299.1/299.95
3	550.33	483.86/476.24/471.39	477.56/482.15/488.38	537.26/538.44/532.58
4	494.79	431.74/428.46/424.3	423.66/428.57/428.4	474.35/479.26/471.19
5	573.25	572.26/592.86/595.12	558.88/562.65/559.71	601.11/593.57/598.84
6	525.99	497.11/478.04/480.39	495.81/468.9/481.7	542.17/540.16/533.21
7	615.5	620.66/573.72/573.83	604.23/586.04/609.4	689.32/685.64/691.15
8	505.24	513.1/484.98/487.69	505.16/485.47/492.52	550.32/542.14/543.32
9	724.36	718.26/694.82/697.24	708.39/679.5/721.29	867.02/846.33/860.86
10	556.03	583.15/575.88/574.17	555.28/523.51/536.77	572.36/563.5/575.2
11	774.3	816.32/784.36/796.39	690.46/654.97/666.67	808.13/785.7/695.75
12	766.93	794.07/745.63/762.01	669.15/662.29/652.06	805.03/774.02/687.7
13	995.67	1027.5/972.31/1001.33	928.88/893.92/881.37	1079.86/1078.04/924.92
14	1117.1	1174.73/1141.99/1127.99	1066.82/1022.62/1029.05	1216.96/1201.9/1066.3

Table 15: **Sys**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-139.85	-241.01/-209.71/-299.36	-248.37/-198.96/-215.09	-173.73/-182.95/-168.51
2	-547.36	-815.15/-745.71/-781.02	-772.25/-661.56/-655.7	-555.37/-562.52/-593.85
3	-207.61	-336.09/-291.69/-330.42	-356.41/-283.4/-320.73	-288.3/-288.0/-270.19
4	-846.97	-910.62/-907.59/-901.25	-835.57/-885.5/-835.86	-791.96/-784.01/-815.54
5	-737.03	-882.27/-942.39/-945.03	-993.23/-933.57/-961.47	-783.31/-775.58/-750.96
6	-1058.66	-1132.95/-1209.46/-1257.06	-1127.49/-1135.7/-1110.98	-1016.27/-974.51/-1058.69
7	-895.81	-1151.67/-1103.31/-1147.42	-1084.43/-1004.98/-1014.51	-961.19/-967.05/-968.65
8	-1261.34	-1448.63/-1467.32/-1419.58	-1419.16/-1434.63/-1425.05	-1297.53/-1305.16/-1300.08
9	-960.95	-1198.73/-1119.61/-1175.89	-1217.99/-1093.91/-1064.58	-990.98/-970.98/-1009.68
10	-1353.94	-1558.4/-1521.83/-1549.88	-1496.75/-1500.04/-1511.18	-1291.28/-1276.94/-1307.21
11	-2835.06	-4111.69/-3688.86/-4030.73	-3864.45/-3625.85/-3816.33	-2731.3/-2815.89/-2832.65
12	-4390.02	-5056.1/-5065.5/-4996.99	-4791.67/-4826.36/-4739.67	-4185.36/-4142.83/-4394.99
13	-2912.41	-3610.54/-3648.31/-3651.87	-3633.98/-3403.51/-3475.73	-2724.99/-2740.86/-3052.02
14	-5120.77	-5888.14/-5785.98/-5703.04	-5746.01/-5675.48/-5731.17	-4989.57/-5120.15/-5107.67

Table 16: **Tam**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-4.76	-32.19/-32.74/-32.51	-51.59/-52.02/-35.45	-47.38/-26.14/-39.83
2	-17.73	-46.54/-47.27/-45.46	-55.7/-56.62/-72.97	-62.05/-38.58/-59.66
3	-16.61	-80.75/-79.8/-78.79	-107.25/-108.55/-71.49	-64.28/-63.6/-54.84
4	-60.94	-110.25/-109.5/-108.27	-119.03/-119.96/-99.67	-105.03/-103.62/-104.16
5	-55.22	-170.33/-172.01/-170.64	-225.89/-225.0/-168.89	-171.4/-158.78/-136.55
6	-79.51	-195.18/-194.07/-193.54	-254.65/-255.64/-181.93	-163.58/-171.47/-162.43
7	-45.79	-220.18/-221.41/-216.07	-251.16/-247.82/-173.37	-180.34/-183.69/-142.69
8	-63.92	-222.53/-222.72/-221.33	-284.38/-284.31/-208.22	-164.19/-233.97/-164.86
9	-21.31	-188.04/-191.03/-187.95	-247.86/-251.34/-132.97	-163.54/-133.88/-111.0
10	-120.25	-350.43/-356.92/-352.06	-460.2/-459.77/-248.28	-235.49/-311.12/-225.59
11	-158.36	-771.48/-762.37/-784.65	-1045.97/-1040.91/-932.19	-626.89/-837.12/-554.16
12	-428.07	-995.03/-972.69/-1001.36	-1152.55/-1148.97/-896.87	-935.4/-999.68/-803.71
13	-340.33	-1069.8/-1062.34/-1088.35	-1672.41/-1675.58/-1032.77	-752.71/-1049.92/-722.25
14	-904.04	-1438.94/-1445.04/-1446.56	-1586.86/-1590.0/-1296.07	-1484.36/-1502.06/-1225.5

Table 17: **Traffic**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	93.18	19.34/11.75/10.37	93.15/93.17/92.97	93.11/92.86/93.22
2	93.97	29.0/26.24/33.14	93.82/93.81/93.82	93.7/93.54/93.83
3	75.7	-31.84/-33.2/-40.0	84.07/82.93/82.82	82.96/82.7/83.0
4	74.2	-24.36/-26.4/-40.0	85.34/84.16/84.36	84.16/84.22/84.27
5	70.56	-38.66/-39.33/-40.0	72.58/72.89/66.05	72.37/72.56/72.93
6	72.11	-39.33/-37.99/-40.0	74.85/75.06/67.26	73.24/65.52/72.06
7	-21.25	-40.0/-40.0/-40.0	15.03/8.92/6.2	15.35/20.17/18.08
8	40.74	-40.0/-40.0/-40.0	45.73/41.63/33.56	45.47/39.38/51.97
9	-40.0	-40.0/-40.0/-40.0	-40.0/-39.49/-39.49	-39.49/-40.0/-40.0
10	-40.0	-40.0/-40.0/-40.0	-38.47/-36.93/-34.86	-36.9/-37.44/-36.4
11	-93.34	-100.0/-100.0/-100.0	41.55/43.69/35.9	44.1/43.46/43.34
12	-62.13	-100.0/-100.0/-100.0	31.32/36.7/27.72	37.44/10.74/34.26
13	-93.79	-100.0/-100.0/-100.0	20.92/24.9/11.66	23.9/24.98/24.88
14	-100.0	-100.0/-100.0/-100.0	13.29/18.56/6.04	18.56/-7.71/8.71

Table 18: **TT**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.

Instance	PROST	SymNet	SymNet-IL	SYMNET2.0
1	-255.5	-5203.7/-5735.5/-5945.6	-700.05/-916.6/-517.3	-581.5/-492.6/-863.7
2	-10073.4	-15285.8/-16088.4/-15243.4	-9062.3/-9258.8/-13719.8	-12324.7/-12507.9/-9989.1
3	-1948.1	-5989.3/-11190.7/-10618.2	-1564.8/-2551.0/-10383.2	-7107.8/-7526.8/-2457.2
4	-22138.1	-12075.3/-20610.2/-21650.8	-9643.9/-10569.0/-23488.0	-17780.3/-15801.8/-10444.4
5	-3071.4	-1366.5/-7736.2/-6614.1	-1526.7/-1052.0/-1384.0	-1235.8/-840.8/-1159.8
6	-16955.6	-22761.0/-26058.1/-23511.9	-8509.4/-7506.9/-26653.2	-16021.6/-16248.9/-8465.0
7	-7901.7	-9600.9/-16326.1/-14335.9	-7409.4/-7764.3/-13439.4	-11142.5/-10848.5/-7460.8
8	-14227.8	-20071.4/-25023.5/-24296.1	-12411.5/-11732.4/-25408.0	-18271.3/-14122.0/-11788.1
9	-13159.4	-12998.1/-17906.7/-16959.9	-10672.7/-13240.9/-17173.6	-14159.6/-14802.5/-11003.5
10	-18557.1	-19537.1/-28020.4/-24454.1	-11849.5/-11287.3/-25618.1	-16793.5/-16994.8/-11161.6
11	-2346.3	-84264.7/-101294.7/-98266.8	-82318.6/-89991.2/-108015.4	-96773.8/-94071.2/-89717.4
12	-1198.7	-82673.6/-109069.7/-101623.7	-52566.75/-50517.1/-113444.5	-86004.7/-86969.9/-50793.7
13	-5331.7	-140856.3/-149905.8/-147387.3	-135853.1/-137694.4/-146235.4	-141409.1/-134314.0/-134457.2
14	-1919.4	-154098.3/-172509.7/-166545.0	-85025.4/-100710.2/-178359.4	-149268.7/-150642.6/-102534.1

Table 19: **Wild**: Table showing raw long term rewards for each neural model averaged over 200 runs in instances 1-10 and 100 runs in instances 11-14. Each entry shows results for 3 runs.