Effectively handling Non-Fluents and Actions (ENFA) in Generalized Neural Policies for RDDL Relational MDPs

Abstract

Relational MDPs (RMDPs) compactly represent an infinite set of MDPs with an unbounded number of objects. Solving an RMDP requires a *generalized* policy that applies to all instances of a domain. Recently, Garg et al. proposed SymNet for this task – it constructs a graph neural network that shares parameters across all instances in a domain, thus making it applicable to any instance in a zero-shot manner. Our analysis of SymNet reveals that it performs no better than random on 1/4th of planning competition domains. The key reasons are its design choices: it misses important information during graph construction, leading to (1) poor generalizability, and (2) potential non-identifiability of different actions.

In response, our solution, ENFANet, substantially augments SymNet's graph construction approach by introducing additional nodes and edges which allow a better transfer of important information about a domain. It also improves SymNet's action decoders with relevant information from objects to make different actions identifiable during scoring. Extensive experiments on twelve competition domains, where we use imitation learning over data generated from the PROST planner, demonstrate that ENFANet performs vastly better than SymNet. Interestingly, even though ENFANet is trained over data from PROST, it outperforms the planner on several test instances due to former's ability to scale to large instances in a zero-shot manner.

1 INTRODUCTION

A Relational Markov Decision Process (RMDP) (Boutilier et al. [2001]) is a first-order representation of a planning domain usually represented in a description language like the Probabilistic Planning Domain Definition Language (PPDDL) [Younes et al., 2005] or the Relational dynamic influence diagram language (RDDL) [Sanner, 2010]. Finding solvers for an RMDP which perform well on any instance of a domain has been a long-standing goal of AI planning research. Motivated by the recent progress in deep neural models, multiple works [Groshev et al., 2018, Toyer et al., 2018, Garg et al., 2019, 2020, Ståhlberg et al., 2021] learn generalized neural reactive policies, which are trained on a set of (smaller) training instances, and can be transferred to a set of (larger) test instances in a zero-shot manner. Our focus is on learning generalized neural policies for RMDPs expressed in RDDL, where SymNet [Garg et al., 2020] has demonstrated initial feasibility.

However, our analysis reveals that SymNet performs no better than random on 1/4th of the domains of the International Probabilistic Planning Competition[1] (IPPC 2011 and 2014), and even in several others where it seemingly does well, it performs significantly worse than PROST [Keller and Eyerich, 2012], the state-of-the-art *online* planner. This points to a significant research gap between what is possible, and what is currently achievable. In this paper, our goal is to examine whether we can fill this gap by a better design of the underlying neural architecture.

At a high level, SymNet compiles an RMDP instance to an *instance graph*, with nodes representing object tuples, and edges representing connections in the Dynamic Bayes Net (DBN) corresponding to the instance. Given a state, a Graph Attention Network [Veličković et al., 2018], on top of the instance graph, computes embeddings for each node. A subset of these nodes embeddings (or their aggregate) is then passed through an action decoder network to output a score for the ground actions. The network is typically trained using a loss function based on reinforcement learning (RL).

We identify two key challenges with SymNet's design choices. First, its handling of *non-fluents*, variables which are static throughout the application of a policy but whose

[1] https://www.icaps-conference.org/competitions/

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

value depends on the given instance, is somewhat ad-hoc. Many non-fluents do not directly correspond to specific nodes in the graph, instead they are compiled away. This leads to a significant problem with generalizability of the network to instances where the value of those non-fluents differs. Second, the action decoder for a ground action takes as input those node embeddings that are affected by the action; it does not necessarily take all the objects that are arguments of the action. This can lead to a problem of action non-identifiability: two ground actions with different object arguments affecting the exact same set of objects get exactly the same score. We describe these in detail through a running example in Section 3.2.

To mitigate these issues we present ENFANet, which substantially augments SymNet's architecture. To handle non-fluents in a principled manner, ENFANet's instance graph creates a node for each object tuple appearing as an argument to any non-fluent. In order to connect these nodes to the rest of the network it additionally creates singleton nodes for each object in the instance. These singleton object nodes connect to all object-tuple nodes that contain this object. To handle action non-identifiability during decoding, we additionally pass the embeddings of all singleton nodes that appear as action arguments in the action.

We train both SymNet and ENFANet with imitation learning on a dataset generated by planning using PROST on training instances; this helps us circumvent the training and exploration issues faced by RL algorithms. Extensive experiments on twelve IPPC domains demonstrate that ENFANet performs vastly better than SymNet, obtaining a gain of more than 40% relative performance on half of the domains, and a gain of approx. 50% relative performance in the aggregate metric. We perform further studies by analyzing specific domains to characterize the various settings in which ENFANet outperforms SymNet. Interestingly, though ENFANet uses data generated from PROST, due to its offline nature, which requires only a forward pass during inference, ENFANet outperforms PROST on large instances of several domains; in some cases by a significant margin. This opens up new avenues for exciting research that combines online planners with policies learned using neural models.

2 BACKGROUND AND RELATED WORK

2.1 RELATIONAL MDPS AND RDDL

A Relational Markov Decision Process (RMDP) [Boutilier et al., 2001] domain, denoted by R_M, represents a factored MDP in a first order form as a tuple $(C, SP, A, \mathcal{O}, T, R, H, s_0, \gamma)$, where SP and A denotes the set of state, respectively, action predicates; $\mathcal{O}$ denotes the set of objects, where each object is associated with a class type in C. The set of transition functions is denoted by T, the set of reward functions by R. Additionally, H denotes the finite horizon and γ the discount factor. Replacing the arguments of a predicate with an object-tuple of type-consistent objects is called grounding the predicate. Grounding the predicates of SP results in a set of state-variables, denoted by $SP_{\mathcal{O}}$, and grounding the predicates of A results in a set of ground actions, denoted by $A_{\mathcal{O}}$. An assignment to all $SP_{\mathcal{O}}$ denotes a state $s \in \mathcal{PS}(SP_{\mathcal{O}})$ where $\mathcal{PS}$ denotes the power set. The initial state is denoted by s_0.

The Relational Dynamic Influence Diagram Language (RDDL) [Sanner, 2010] represents an RMDP using two components: 1) a domain description provides predicates SP and A, object types C, as well as first-order transition and reward functions T and R; and 2) an instance description specifies ground objects $\mathcal{O}$, initial state s_0, as well as horizon H and discount factor γ. Furthermore, the set of state predicates (SP) is divided into state-fluents (SF) and non-fluents (NF), where the former are predicates where the assignment of induced ground variables can change over time, and the latter are predicates whose ground variables' assignment remains static. Note that two instances induced by the same domain can have different assignments of ground variables induced by NF. We denote with O_{SF} and O_{NF} the set of object tuples that appear in SF, respectively NF. Given a RDDL instance, its transition semantics can be represented in the form of a Dynamic Bayesian Network (DBN) capturing dependencies among state-variables and ground actions [Sanner, 2010].

2.2 TRANSFER LEARNING FOR RMDPS

We define the problem of *Transfer Learning for RMDPs (TLR)* as follows. Given an RMDP R_M and a set of instances of R_M expressed in RDDL, the goal of TLR is to learn a generalized neural network $\mathcal{N}(I)$ parameterized by instance I, with a (tied) set of weight parameters w independent of I, such that $\mathcal{N}(I)$ takes as input a state s of instance I, and outputs a distribution over actions in the action space of I, i.e. $\mathcal{N}(I) : \mathcal{PS}(SP_O) \rightarrow p(A_O)$ where $p(A_O)$ represents a probability distribution over all ground actions A_O .

2.3 RELATED APPROACHES

Generalized planning has long history with early work trying to construct features which transfer across instances Fern et al. [2003], Guestrin et al. [2003]. Recent work uses the notion of fully observable non-deterministic planners (FONDS) for generalized planning Bonet and Geffner [2018], Bonet et al. [2019]; all these works are non-neural in nature. There is some recent work Toyer et al. [2018] which builds neural models over PPDDL, but since our focus is on RMDPs expressed in RDDL, and the architecture of neural reactive policies is tailored to the description language, these works are not directly comparable to ours. Issakkimuthu et al. [2018] learn Deep Reactive Policies for

RDDL domains, however, their model is not capable of size transfer. We instead build upon a series of works [Bajpai et al., 2018, Garg et al., 2019, 2020] which proposes neural solvers for RDDL. Torpido [Bajpai et al., 2018] can only perform transfer on instances of same size, whereas TrapsNet [Garg et al., 2019] makes additional assumptions on the arities of state and action predicates. Closest to our work is SymNet (Garg et al. [2020]), which, to our knowledge, is the only neural model for a general RDDL RMDP. We now describe its architecture in detail.

2.4 SYMNET

Given an RDDL domain and an instance I, SymNet (Garg et al. [2020]) solves TLR as follows: 1) first, represent I in the form of an *instance-graph*, 2) use a GAT-based architecture to represent the generalized policy, 3) finally, train the model using a suitable end-to-end loss, e.g. RL-based or imitation learning based - we compare with both in our experiments. Next, we will discuss these steps in detail.

Instance-Graph Construction: We start by discussing how SymNet creates its instance-graph. In SymNet, the purpose of the instance-graph(s) is to translate an instance into graph(s) that capture interactions among various state-variables. For this, SymNet creates $|A| + 1$ graphs, $\mathcal{G}_{sym} = \{G_d, G_{a1}, \ldots, G_{a|A|}\}$. All graphs are derived from the DBN of the instance: G_d captures exogenous, i.e. action-independent effects between state-variables, and each $G_{ai} \in \{G_{a1}, \ldots, G_{a|A|}\}$ captures effects between state-variables that are induced by action ai.

Recall that O_{SF} represents the set of object tuples that appear in state-fluents. For each $o_{sf} \in O_{SF}$ SymNet adds a node v with label o_{sf} to each of the $|A| + 1$ graphs. Edges are introduced once all nodes are generated. In the following, let v_1 and v_2 be two nodes labeled with object tuples o_1, respectively, o_2. Whether an edge exists between v_1 and v_2 depends on the underlying graph:

1) for G_d there is an edge between v_1 and v_2 if the DBN contains a state-variable $SP(o_1)$ that affects another state-variable $SP(o_2)$. Note that every state-variable affects itself, hence every node has a self-loop.

2) for graph $G_{ai} \in \{G_{a1}, \ldots, G_{a|A|}\}$ there exists an edge between v_1 and v_2 if there is a state-variable $SP(o_1)$ and an action $a(o_a) \in A_{\mathcal{O}}$ of type $ai \in A$, that in conjunction affect another state-variable $SP(o_2)$. That is, it captures if a state-variable and some action of type ai affect some other state-variable in the DBN.

Node Features: All graphs have the same set of input node features, determined by the following rules: a) For each parameterized predicate type $P \in SF$, a feature is added to every node v. For each grounding $P(o)$, the node feature of o that corresponds to P is set to the value of $P(o)$. The value is fetched from the current state. b) For each unparameterized Boolean non-fluent, a feature with its value is added to each node. c) A feature for a parameterized Boolean non-fluent is added to a node, if the object tuple corresponding to the non-fluent is a subset of the object-tuple at the node.

Node Embeddings: SymNet uses a Graph Attention Network (GAT) [Veličković et al., 2018], which is a specific kind of graph neural network that leverages the attention mechanism over a node's neighbors for its message passing updates. SymNet uses a GAT to compute node embeddings for each graph in $\mathcal{G}_{sym}$. We establish a correspondence between nodes in different graphs having the same label, i.e., which correspond to the same object tuple. A final node embedding $ne(v)$ for a node v (representing all the nodes in different graphs having the same label) is constructed by: $ne(v) = concat(GAT_d(G_d)[v], ..., GAT_{a|A|}(G_{a|A|})[v])$. A global embedding ge representing the complete state is then computed as a maxpool over all node embeddings as: $ge = maxpool_{v \in V}(ne(v))$ where V is the set of all nodes.

Action Decoding: SymNet creates a set of action decoders $(AD_1, \ldots, AD_{|A|})$ for each action type in the domain. Let there be a parameterized ground action $a(o)$ that affects a set of state-variables $\mathcal{P}_{a(o)}$. Let $args(P)$ denote a function that returns the arguments of predicate P. Then, the score of action $a(o)$ is computed as $score(a(0)) = AD_{type(a)}\big(maxpool_{P \in \mathcal{P}_{a(o)}}(ne(args(P))), ge\big)$, where $type(a)$ returns the type of action a. To get a policy, $softmax$ is taken over all action scores.

3 ENFANET: A NEW ARCHITECTURE

We formally discuss the shortcoming of SymNet's instance-graph and its architecture. We then propose **E**ffectively handling **N**on-Fluents and **A**ctions **N**etwork (ENFANet) to learn a generalized neural policy $\mathcal{N}(I)$ for zero shot transfer.

3.1 RUNNING EXAMPLE

Recon is an IPPC domain where the agent moves in a 2D grid-world and is equipped with tools for detecting water, life, and taking pictures. Certain locations on the grid are marked as hazard and if the agent uses a tool on these locations the tool gets damaged with a high probability. Once a tool is damaged the agent has to return to the base location where they can repair the tool. The agent is positively rewarded for taking pictures of cells where life is detected. The domain has:

Objects Types: `x,y,obj,agent,tool`.

Non-Fluents: `objAt(obj, x, y)`, `is_up(y1,y2)`, `is_down(y1,y2)`, `is_right(x1,x2)`, `is_left(x1,x2)`, `base(x, y)`, `hazard(x, y)` , `detect_prob_damaged`, `damage_prob(tool)`, `detect_prob`, `camera_tool(tool)`,

life_tool(tool), water_tool(tool), good_pic_weight, bad_pic_weight.
State-Fluents: agentAt(agent, x, y), damaged(tool), waterChecked(obj), waterDetected(obj), lifeChecked(obj), lifeChecked2(obj), lifeDetected(obj), picTaken(obj).
Actions: up(agent), down(agent), left(agent), right(agent), useToolOn(agent, tool, obj), repair(agent, tool)

We consider an instance with a 2×2 grid, where $\{x_1, x_2\}$ and $\{y_1, y_2\}$ are of type x, respectively y. There is a single agent ag_1, two tools $\{t_1, t_2\}$ and two objects $\{o_1, o_2\}$.

3.2 SHORTCOMINGS IN SYMNET

As motivated in Section 1, SymNet makes certain design choices which results in sub-optimal performance on several planning problems. First, since its instance graph is derived from the underlying DBN, it is incapable of capturing important information present in the RDDL description in the form of parameterized non-fluents. Specifically, SymNet's instance graph can only incorporate information about those non-fluents whose arguments also appear in a state-fluent; for all others, the information is compiled away. Second, the score of each action is decided solely on the basis of what state-variables the action affects. This means that any action arguments which do not appear in state-fluents affected by the action will have no impact on the action score, resulting in action non-identifiability as demonstrated by the following proposition. Given an action $a(o)$, we will use the notation $\mathcal{P}_{a(o)}$ to denote the set of state-variables (fluents) affected by $a(o)$.

Proposition 1. *Let there be two actions $a(o_1)$ and $a(o_2)$ of action type $type(a)$, where $o_1 \neq o_2$. Let both actions affect the 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 identical.* [see Appendix for a proof]

In our example, non-fluent objAt(obj,x,y) indicates that the object obj is present at the location x,y, but since there is no state-fluent with this set of arguments, the grounding of this object tuple is never represented explicitly in the instance graph. Hence, the network may not generalize well to instances where objects are present at different locations than those seen during training. Further, there is an action useToolOn(agent,tool,obj) which says that agent uses tool on obj. Since this action only affects state fluents with object tuple obj, the embedding for tool is not incorporated during action decoding, resulting in an identical score for two actions applying different tools to the same object.

Because of above issues, SymNet results in learning sub-optimal policies which do not transfer well to new instances for several domains. Next, we describe our approach which can handle these shortcomings in a comprehensive manner.

3.3 OUR APPROACH

To handle these shortcomings we will make two changes, 1) we add a set of new graphs to SymNet, and 2) we add new inputs to the action decoder. We explain these details next.

Adding Position-based Graphs: On top of graphs in SymNet, we create a new set of graphs $\{G_{p1}, \ldots, G_{p|Ar|}\}$ that capture what object comes at what position in a state-variable or non-fluent. Hence, we now have $\mathcal{G}_{enfa} = \{G_d, G_{a1} \ldots, G_{a|A|}, G_{p1}, \ldots, G_{p|Ar|}\}$, where $|Ar|$ is the maximum arity of any predicate in the domain.

Intuitively, these new graphs capture the relationship between object tuples in the instance, which could be part of a state-fluent or a non-fluent, and their individual object arguments. There is a different graph for each position that an argument could appear in, in order to capture the relative ordering of arguments. We next describe the set of nodes and edges in these newly added graphs:

1) **Object Tuple Nodes:** For each $o_{sf} \in O_{SF}$ we add a vertex u to each of the $|Ar|$ graphs with label o_{sf}. Note that these nodes are the same as those in SymNet's instance-graph. Similarly, for each $o_{nf} \in O_{NF}$ we add a vertex v to each of the graphs with label o_{nf}. [2] These nodes are added to capture the missing information available in non-fluents which is not covered by SymNet.

2) **Singleton Object Nodes:** Finally, for each $\tilde{o} \in \mathcal{O}$ a vertex w with label $\tilde{o}$ is added to the graphs (if it is not already added in the previous step). These new singleton object nodes are created for message passing to and from non-fluent based nodes. As a side benefit, we will see later that these singleton object nodes will also be helpful in removing action non-identifiability.

For each object-tuple $o \in O_{SF} \cup O_{NF}$, and for each object $o[i] \in \mathcal{O}$ appearing at position i in o, we add edges $e(o, o[i])$ and $e(o[i], o)$ in G_{pi}. This means, each graph in $\{G_{p1}, \ldots, G_{p|Ar|}\}$ has bidirectional edges that capture whether an object occurs at position i of any object-tuple (of any state-variable or non-fluent). Here, separate adjacency for each position is used to preserve ordering of objects in an object-tuple. This helps in preserving semantic meaning in predicates like is_up(a,b) where ordering of a and b matters, hence, is_up(a,b) and is_up(b,a) should be treated differently.

Node Features: All newly constructed graphs have the same set of input node features, which are described as follows:

[2] In order to be memory efficient, we add these nodes only for non-fluents taking non-default value.

1) **State-Fluent Features:** For each parameterized state predicate type P, we add a feature to every node v. For each grounding $P(o)$ of P, the node feature of o that corresponds to P is set to the value of $P(o)$ fetched from the current state. For all other object tuples which do not appear as groundings of P this feature is set to null. We denote the set of the resulting features with $h^{SF}(v)$.

2) **Non-Fluent Features:** For each parameterized non-fluent predicate type P, we add a feature to every node v. For each grounding $N(o)$ of N, the node feature of o that corresponds to N is set to the value of $N(o)$. The value is fetched from the instance description for the latter, and from the domain description for the former. For all other object tuples which do not appear as groundings of P this feature is set to null. We denote the set of the resulting features with $h^{NF}(v)$.

3) **Global Features:** Unparameterized state-fluents and non-fluents represent global properties relevant to all nodes, hence, these are added as features to every node. The values are fetched from the current state for state-fluents and from the instance description for non-fluents. Let these features be denoted by h^G.

4) **Type Features:** For each node v with label o, we create a one-hot encoding vector $h^{TY}(v)$ representing the type of the node in the instance-graph(s). We define the type of each object-tuple $o = (o_1, \ldots, o_l)$ as $type(o) = (type(o_1), \ldots, type(o_l))$ where the type operator is overloaded to return the type of object given as input to it.

The overall node feature of a node v is represented as: $h(v) = concat(h^{SF}(v), h^{NF}(v), h^G(v), h^{TY}(v))$.

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 $d_{sym}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in $\mathcal{G}_{sym}$ and let $d_{enfa}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in $\mathcal{G}_{enfa}$. Then, $d_{enfa}(u, v) \leq d_{sym}(u, v)$.* [see Appendix for a proof]

Proposition 2 shows that $\mathcal{G}_{enfa}$ can have shorter distances among nodes in the graph. This can result in better message passing as also demonstrated in Section 4.2.3.

Node Embedding: We use the similar GAT-based architecture as in SymNet to compute node embeddings for each graph in $\mathcal{G}_{enfa}$. Like in SymNet, we establish a correspondence between nodes in different graphs having the same label, i.e., which correspond to the same object tuple. A final node embedding $ne(v)$ for a node v (representing all the nodes in different graphs having the same label) is constructed by: $ne(v) = concat(GAT_d(G_d)[v], \ldots, GAT_{a|A|}(G_{a|A|})[v], \ldots, GAT_{p|Ar|}(G_{p|Ar|})[v])$. In case there is no corresponding node in a graph, it must be present in the newly added graphs, and we use the corresponding input features. To represent the complete state, a global embedding ge is then computed as a maxpool over all node embeddings as: $ge = maxpool_{v \in V}(ne(v))$, V being the set of all nodes.

Action Decoding: To address the issue with SymNet's decoding, while computing the score of a parameterized action $a(o)$, we also give as input the node embeddings of each object occurring as a parameter in $a(o)$ along with the node embeddings of the nodes it affects. This leads to unique identification of each action as its parameters uniquely identify it. Formally, let there be a parameterized ground action $a(o)$ that affects a set of state-variables $\mathcal{P}_{a(o)}$ and let $o = (o_1, \ldots, o_n)$ then, the score $score(a(o))$ is given as: $AD_{type(a)}\big(ne(o_1), \ldots, ne(o_n), maxpool_{P \in \mathcal{P}_{a(o)}}(ne(args(P))), ge\big)$.

3.4 TRAINING ALGORITHM

We use a two phase process to train ENFANet using imitation learning. In the first phase, referred to as dataset generation, for each training instance in the set of training instances I_{tr} we use the PROST [Keller and Eyerich, 2012] planner, a state-of-the-art UCT-based *online* probabilistic planner, to generate a set of trajectories $\tau_1, \ldots, \tau_M$, where each trajectory is a sequence of state-action pairs $\langle s_0, a_0, \ldots, s_{H-1}, a_{H-1} \rangle$. To compute dataset D_i we first compute the union of all state-action pairs among all trajectories. Since PROST is a sampling-based planner with time-limited lookahead, different trajectories can potentially contain state-action pairs (s, a_i) and (s, a_j), i.e. pairs which share the same state, but where a different action is applied. This may cause problems for the underlying neural learner.We circumvent this problem by only keeping the action which occurs most frequently for a given state and leave the exploration of other solutions as a direction for future work.

In the second phase, referred to as neural learning, ENFANet is trained using supervised learning using the dataset generated in Phase 1 above. During training, we divide each D_i into batches and we consume all batches of D_i before moving to the dataset of the next instance. A cross-entropy based loss is used during training. During inference we take an $argmax$ over the action distribution to decide the action to be taken. We note that in the work done by Garg et al. [2020], SymNet was trained using an RL based loss. For a fair comparison, we experiment with SymNet using both kinds of losses, i.e., an RL based loss and imitation learning based loss, as described above.

4 EXPERIMENTS

With our experiments, we want to answer three key questions. (1) IPPC performance: does ENFANet result in better performance on IPPC instances compared to SymNet? (2) how well does ENFANet generalize to instances that go

far beyond the size of the largest IPPC instances, compared to other approaches? (3) how well does ENFANet generalize to instances where there is a significant difference between the non-fluents of the test instance and the non-fluents seen during training?

4.1 EXPERIMENTAL SETUP

Domains: We evaluate all models on twelve IPPC 2011 and 2014 domains: Academic Advising (Acad), Crossing Traffic (CT), Game of Life (GoL), Navigation (Nav), Skill Teaching (Skill), Sysadmin (Sys), Tamarisk (Tam), Traffic, Wildfire (Wild), Recon, Triangle Tireworld (TT) and Elevators (Elev) (see the domain descriptions in the Appendix). For each domain, we pick IPPC instances 1-3 as training instances, do validation on instance 4 and test on instances 5-10 (unless stated otherwise for some experiment). While training any model, we save checkpoints at regular intervals. We validate on instance 4 by evaluating each saved checkpoint and pick the one with the best reward for final testing.

Algorithms & Settings: SymNet is the only published architecture for the task of training a generalized neural policy for RDDL RMDPs. It uses RL to train, which, in our preliminary experiments, suffers from exploration issues, due to the sparse rewards inherent to many IPPC domains. Since ENFANet is trained using imitation learning (IL), we create a stronger baseline by training the SymNet architecture also with the IL data. We name this system SymNet-IL. To construct IL data, for each training instance, we run PROST[3] in its default setting and collect 100 trajectories, which are converted to (state, action) pairs and used as IL training data in a domain.

SymNet is trained for 12 hours (as per original paper's setting). ENFANet and SymNet-IL are trained for 500 epochs with a maximum allowed training time of 12 hours (for parity). However, in practice, both IL-based models are much faster to train and take no more than 7 hours training (including data generation) in any domain.

We are guided by the literature on domain independent planning, where the goal is to develop a *single* planner that can work on any domain. So, we do not apply any domain specific hyperparameter tuning, and use a fixed neighborhood size of 1 in the GAT for all domains. Section 4.2.1 briefly discusses the effect of this hyperparameter.

Finally, we also compare against PROST. We emphasize that any direct comparison with PROST is not meaningful, as PROST is an online planner that uses interleaved planning and execution and the other three models are offline planners. Nevertheless, we still include this comparison to gain a deeper insight into our results (generally, the expectation is that PROST will perform better as it can target interleaved exploration for the states that are actually reached). On test instances, PROST is evaluated in its default setting. This implies that at test time it will be slower than the other approaches, but its overall training plus test time will be lower.

Evaluation Metric: We follow existing literature on neural MDP solvers [Bajpai et al., 2018, Garg et al., 2019, 2020] and use the evaluation metric (α) that outputs a number between 0 and 1, with 0 denoting a performance equal to random, and 1 denoting the best reward amongst all comparison approaches. In more detail, for a given domain, we run the train-validate-test cycle 3 times for each model m (neural models, PROST, and random policy). For the rth run of m, we execute its policy for 200 episodes on each test instance i, and store the average long term reward as $V(m, i, r)$. The maximum value of $V(m, i, r)$ is denoted as $V_{max}(i)$, and $V_{rand}(i)$ is the long term reward of the random policy.

Next, we assess the relative performance of a policy by computing a normalized metric $\alpha(m, i, r) = \frac{V(m,i,r)-V_{rand}(i)}{V_{max}(i)-V_{rand}(i)}$. To estimate the performance of a model m on a domain, we compute $\alpha(m) = \frac{1}{|r|}\sum_r \frac{1}{|i|}\sum_i \alpha(m, i, r)$. If this metric is 1, that means that it outputs the best score in every instance. A negative value denotes that it outputs worse than random policies on average.

4.2 RESULTS

Table 1 reports our main result – all models tested on 12 IPPC domains in the setting described above. Each $(m, d)^{th}$ entry represents $\alpha(m)$: the performance of algorithm m on domain d. The last column shows the mean over all 12 domains. Results of PROST are in gray color, as those numbers are not suitable for a direct comparison, but give a deeper insight into the overall performance quality. The bold values show the neural model with maximum $\alpha(m)$.

Overall, ENFANet outperforms SymNet-IL and RL based SymNet by vast margins of $+22$ and $+36$ points, respectively. In particular, ENFANet is better than the improved baseline SymNet-IL in 10 out 12 IPPC domains, and very close in the eleventh (TT). SymNet-IL gets superior results compared to SymNet, underscoring the difficulty in RL based training, and the value of imitation learning. Another noteworthy point is that in no domain is ENFANet's performance close to or worse than random (see Recon and Skill for comparison with SymNet-IL), suggesting that the new instance graph with a better treatment of non-fluents improves the overall model generalization.

4.2.1 Ablation on Neighborhood Size

We determing the influence of the hyperparameter, neighborhood size of the GAT, by varying it from 1 to 3. For both

[3] https://github.com/prost-planner/prost

IPPC Test Instances 5-10													
Model	**TT**	**CT**	**Acad**	**Elev**	**Tam**	**Nav**	**GoL**	**Skill**	**Sys**	**Wild**	**Traffic**	**Recon**	**Mean**
r_1 PROST	0.53	0.86	0.47	1.00	0.94	0.88	1.00	1.00	0.65	0.70	1.00	0.99	0.84
r_2 SymNet	0.00	0.37	0.58	0.31	0.55	0.53	0.20	-0.40	0.62	0.27	0.00	0.03	0.26
r_3 SymNet-IL	**0.83**	0.91	0.72	0.38	0.63	**0.56**	0.20	-0.50	0.49	0.72	-0.18	0.03	0.40
r_4 ENFANet	0.81	**0.95**	**0.82**	**0.44**	**0.92**	0.47	**0.29**	**0.43**	**0.94**	**0.77**	**0.28**	**0.30**	**0.62**
Larger Instances													
Model	**TT**	**CT**	**Acad**	**Elev**	**Tam**	**Nav**	**GoL**	**Skill**	**Sys**	**Wild**	**Traffic**	**Recon**	**Mean**
r_5 PROST	0.09	0.55	0.39	1.00	0.90	0.44	0.91	1.00	0.36	1.00	1.00	0.78	0.70
r_6 SymNet	0.00	0.14	0.60	0.15	0.43	0.41	0.60	-0.82	**0.51**	0.09	0.25	0.02	0.20
r_7 SymNet-IL	**0.96**	0.62	0.63	**0.22**	0.52	0.19	0.25	-0.79	-0.65	**0.22**	0.03	0.02	0.19
r_8 ENFANet	0.95	**0.89**	**0.77**	0.19	**0.94**	**0.95**	**0.84**	**0.34**	0.46	0.20	**0.39**	**0.32**	**0.60**

Table 1: Results showing comparison between ENFANet and the baselines on 12 IPPC domains. All models are trained on (smaller) instances 1-3 and validated on instance 4. Rows r_1 to r_4 show results on IPPC test instances 5-10 and rows r_5 to r_8 show results on much larger instances than those in the IPPC. Bold values show the best performer among all neural models.

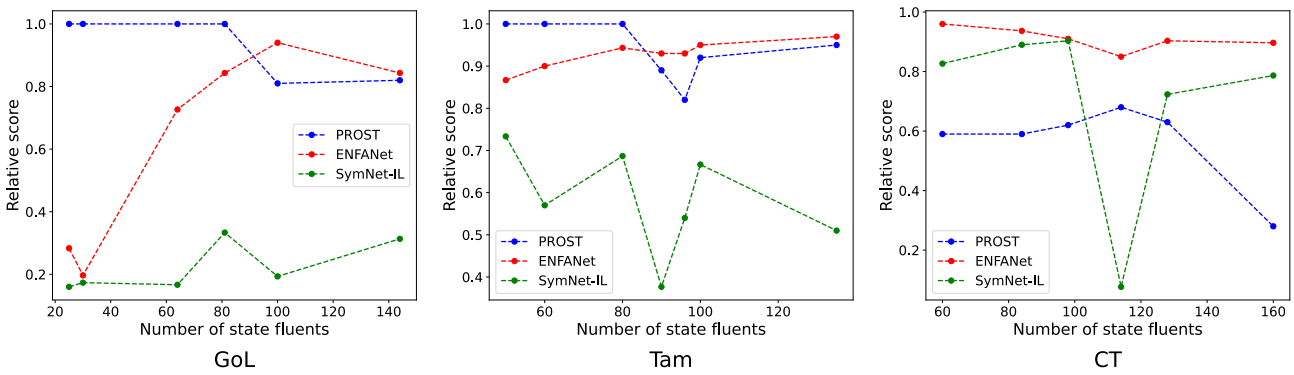

Figure 1: Performance trends on instances of increasing size: PROST deteriorates, but ENFANet remains robust.

Symnet-IL and ENFANet, increasing the neighborhood size to 2 increases the performance in some domains (TT, Acad, Elev, Skill and Recon), but decreases performance in others, causing an overall decrease in performance. For best performance on a domain, this hyperparameter tuning could be easily done on the validation instance. Detailed results are available in the Appendix in Table 2. For rest of the section, unless otherwise stated, we set this parameter to 1.

4.2.2 Offline vs. Online Planning on Larger Instances

When comparing results of the online planner (PROST) with ENFANet, we find that, overall, generalized neural policies are not able to match up to interleaved planning and execution. This is not entirely surprising, since the latter can target exploration based on specific observed outcomes of actions taken earlier. However, interestingly, we find that in a few domains (e.g., TT, Acad), ENFANet is able to outperform PROST. We hypothesize that this could be due to ENFANet's ability to generalize well to large instances.

To test this hypothesis, we create four new test instances[4] for each domain (we call them instances 11 to 14), with sizes much larger than IPPC instances.[5] For some of the domains our instance#14 has three times the number of objects of IPPC's instance#10. For example, TT instance#10 has 66 grid cells, where our instance#14 has 190. Similarly, Acad instance#10 has 30 courses, where our instance#14 has 90. See Appendix A.4 for details on exact sizes. Additionally, we increase the horizon to 100 for these larger instances.

Table 1 shows the comparison. We first notice that the gap between SymNet-IL and ENFANet increases drastically, when tested on larger instances (compared to previous experimental setting). This suggests that ENFANet generalizes more robustly to large problem sizes. We then compare the same gaps between PROST and ENFANet, and find that, in aggregate, ENFANet closes in on PROST, and reduces the performance gap. In 8 of 12 domains (TT, CT, Acad, Tam,

[4] We will release these instance files for further research.

[5] generated using the official scripts provided by the IPPC at https://github.com/ssanner/rddlsim

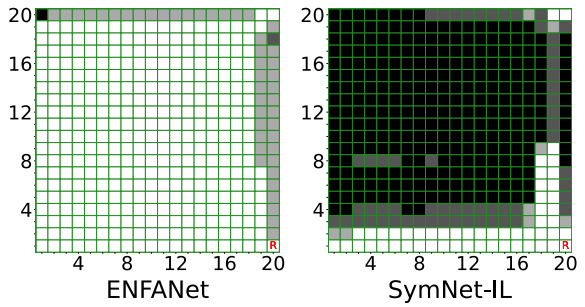

Figure 2: Coverage of ENFANet (left) and SymNet-IL (right) on grid size 20×20 when trained on grid size 5×5.

Nav, GoL, Traffic, Recon) the gap is reduced, whereas it gets worse in only 4 domains.

Figure 1 shows that PROST's relative performance starts to drop, as size increases. Two interesting cases are GoL and Tam, where in aggregate ENFANet performs worse than PROST, but in the figure, we observe that for the largest instances (13 and 14), it starts to outperform PROST. We conjecture that the reason for such results is that larger instances have larger state spaces, branching factors and reward horizon, due to which UCT based online planners like PROST may struggle to find high reward trajectories. In such scenarios, the size-invariance of generalized neural policies makes their additional benefit even more evident.

4.2.3 Generalization to Changing Non-fluents

Non-fluents of a domain control the underlying structure and parameters that affect the transition model and are critical for finding a good policy for a given instance. The non-fluent values vary from instance to instance, and hence it is important for a generalized policy to be robust to these changes. In most IPPC domains, these non-fluents vary considerably and hence our results in Table 1 already provide some evidence for our model's ability to adapt. However, we hypothesize that the gains should not be attributed only to a better non-fluent handling, but also to the newly added singleton nodes. We believe that these singletons facilitate better localization and sharing of information.

To verify this, we create a simple variation of the Navigation domain (without action stochasticity) and vary the goal non-fluents. Similar to a regular Navigation domain, the robot always starts at the lower right corner of a 2D-grid and has to reach a goal using five actions: North, South, East, West and *noop*. It gets a reward of 0 on reaching the goal and -1 otherwise. A state-fluent `robotAt(x,y)` and a non-fluent `goalAt(x,y)` specify the locations of robot and goal respectively. In IPPC instances, the goal non-fluent is always at the upper right corner. However, in our experiment, we test the model by marking each grid cell as the goal in turn – essentially checking the model's ability to learn to solve simple path planning problems.

We train SymNet-IL and ENFANet on instances of size 5×5. To factor out any lack of diversity in training data, we create 24 training instances, one for each grid cell as a goal. For validation we create three instances of size 11×11 where the goal is kept at locations $(4, 4)$, $(4, 7)$, and $(5, 5)$ (see Figure 2 for reference) and the model with the best average reward on these is selected. For testing, a total of 399 instances of size 20×20 are used. Here, we use a GAT neighborhood size of 2 for both models, as it works better for this domain. To gauge the performance of SymNet-IL and ENFANet, we report the test instances where the robot is able to reach the goal. Figure 2 shows the coverage of the robot on the test instances averaged over three runs for both models, where each cell has one of the four colors: black, dark grey, light grey and white, denoting the coverage ratios of 0/3, 1/3, 2/3 and 3/3, respectively, for instances where the goal is located at that cell. We see that the coverage for ENFANet is enormously higher than for SymNet-IL.

Further analysis reveals that the instance graphs of both models already incorporate the knowledge of `goal(x,y)` as a feature in node `(x,y)`. Hence, the better coverage of ENFANet cannot be due to a better handling of non-fluents. The main difference in the two graphs is the addition of singleton nodes and corresponding edges between object tuple nodes `(x,y)` in the position based graphs in $\mathcal{G}_{enfa}$. We believe that these singleton nodes lead to better information exchange among nodes. Nodes `x` and `y` can act as representatives of rows and columns: if the goal is at location `(x,y)`, then the node `x` could learn features like `robotAt(x,*) ∧ goalAt(x,*)` (`*` represents don't care), i.e., a feature that signifies whether the robot is in the same column (analogously row) as the goal. In case of SymNet, singleton nodes are absent, hence it requires message passing steps of arbitrary length to localize the goal, thus, hurting its generalizability.

5 CONCLUSION

We present ENFANet, a neural architecture for learning generalized policies for relational MDP domains expressed in RDDL. Its key technical contribution is a better handling of non-fluents by creating nodes for object tuples that occur as arguments to a non-fluent. It also creates singleton object nodes, when not present, and uses these in the action decoder, which mitigates the problem of action non-identifiability in the previous SymNet system. Extensive experiments reveal that not only is ENFANet vastly superior to SymNet, it is also more robust to large instance sizes, and generalizes well with changing non-fluents. Directions for future work include combining PROST with ENFANet, and extending it to other settings such as Concurrent MDPs [Mausam and Weld, 2004] and POMDPs.

References

Aniket Bajpai, Sankalp Garg, and Mausam. Transfer of deep reactive policies for MDP planning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 10988–10998, 2018.

Blai Bonet and Hector Geffner. Features, projections, and representation change for generalized planning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Blai Bonet, Guillem Frances, and Hector Geffner. Learning features and abstract actions for computing generalized plans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2703–2710, 2019.

Craig Boutilier, Raymond Reiter, and Bob Price. Symbolic dynamic programming for first-order mdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pages 690–700, 2001.

Alan Fern, Sungwook Yoon, and Robert Givan. Approximate policy iteration with a policy language bias. *Advances in neural information processing systems*, 16, 2003.

Sankalp Garg, Aniket Bajpai, and Mausam. Size independent neural transfer for RDDL planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 631–636, 2019.

Sankalp Garg, Aniket Bajpai, and Mausam. Symbolic network: generalized neural policies for relational mdps. In *International Conference on Machine Learning*, pages 3397–3407, 2020.

Edward Groshev, Aviv Tamar, Maxwell Goldstein, Siddharth Srivastava, and Pieter Abbeel. Learning generalized reactive policies using deep neural networks. In *2018 AAAI Spring Symposium Series*, 2018.

Carlos Guestrin, Daphne Koller, Chris Gearhart, and Neal Kanodia. Generalizing plans to new environments in relational mdps. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1003–1010, 2003.

Murugeswari Issakkimuthu, Alan Fern, and Prasad Tadepalli. Training deep reactive policies for probabilistic planning problems. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, 2018.

Thomas Keller and Patrick Eyerich. Prost: Probabilistic planning based on uct. In *Twenty-Second International Conference on Automated Planning and Scheduling*, 2012.

Mausam and Daniel S. Weld. Solving concurrent markov decision processes. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, pages 716–722, 2004.

Scott Sanner. Relational dynamic influence diagram language (rddl): Language description. *Unpublished ms. Australian National University*, 32:27, 2010.

Simon Ståhlberg, Blai Bonet, and Hector Geffner. Learning general optimal policies with graph neural networks: Expressive power, transparency, and limits. *arXiv preprint arXiv:2109.10129*, 2021.

Sam Toyer, Felipe Trevizan, Sylvie Thiébaux, and Lexing Xie. Action schema networks: Generalised policies with deep learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Håkan LS Younes, Michael L Littman, David Weissman, and John Asmuth. The first probabilistic track of the international planning competition. *Journal of Artificial Intelligence Research*, 24:851–887, 2005.

A APPENDIX: IMITATION LEARNING BASED SOLVER FOR RELATIONAL MDPS IN RDDL

A.1 PROOFS OF PROPOSITIONS

Proposition 1. *Let there be two actions $a(o_1)$ and $a(o_2)$ of same action type $type(a)$ 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, $score(a(o_1)) = AD_{type(a)}\big(maxpool_{P \in \mathcal{P}_{a(o_1)}}(ne(args(P))), ge\big) = score(a(o_2))$ 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 $d_{sym}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in $\mathcal{G}_{sym}$ and let $d_{enfa}(u, v)$ denote the minimum distance between nodes u and v in any of the graphs in $\mathcal{G}_{enfa}$. Then, $d_{enfa}(u, v) \leq d_{sym}(u, v)$.*

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 $\mathcal{G}_{enfa}$ because by construction $\mathcal{G}_{enfa}$ contains all nodes and edges of $\mathcal{G}_{sym}$. Hence, $d_{enfa}(u, v)$ is at least equal to $d_{enfa}(u, v)$. 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 $d_{enfa}(u, v) = min(l, 2) \leq d_{sym}(u, v)$.

A.2 ARCHITECTURE DETAILS

ENFANet 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 $\bar{v} = mlp(concat(GAT_1(G_1), ..., GAT_{|N|}(G_{|N|})))$ 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 $\bar{s} = maxpool(\bar{v})$. 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

1. **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.

2. **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.

3. **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 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

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
ENFANet	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
ENFANet	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
ENFANet	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 ENFANet and the baselines on 12 IPPC domains when we vary the neighborhood size of GAT (column "GAT").

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 LARGE INSTANCE GENERATION

We generate 4 instances of increasing sizes for all 12 domains. For this, we use generators provided by official Repository of RDDL Simulator by Scott Sanner [6]. 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.

[6] https://github.com/ssanner/rddlsim/tree/master/src/rddl/competition/generators

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.5 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.6 RAW NUMBERS AND OTHER RESULTS

	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 ENFANet	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 ENFANet	**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 ENFANet 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 ± 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 ENFANet	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 ENFANet	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 ENFANet 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±0.00	0.14±0.01	0.60±0.05	0.15±0.02	0.43±0.02	0.41±0.19
r_3 SymNet-IL	**0.96±0.02**	0.62±0.05	0.63±0.10	**0.22±0.05**	0.52±0.01	0.19±0.01
r_4 ENFANet	0.95±0.03	**0.89±0.08**	**0.77±0.07**	0.19±0.03	**0.94±0.03**	**0.95±0.02**
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±0.01	-0.82±0.02	**0.51±0.10**	0.09±0.03	0.25±0.00	0.02±0.00
r_3 SymNet-IL	0.25±0.17	-0.79±0.00	-0.65±0.08	**0.22±0.09**	0.03±0.10	0.02±0.00
r_4 ENFANet	**0.84±0.03**	**0.34±0.13**	0.46±0.35	0.20±0.04	**0.39±0.08**	**0.32±0.08**

Table 6: Results showing comparison between ENFANet 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 ± 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 ENFANet	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 ENFANet	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 ENFANet 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**	**ENFANet**
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	ENFANet
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	ENFANet
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: **Elev**: 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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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	ENFANet
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.