Supplementary Material: Submission #316

8.1 IPC Domains Statistics

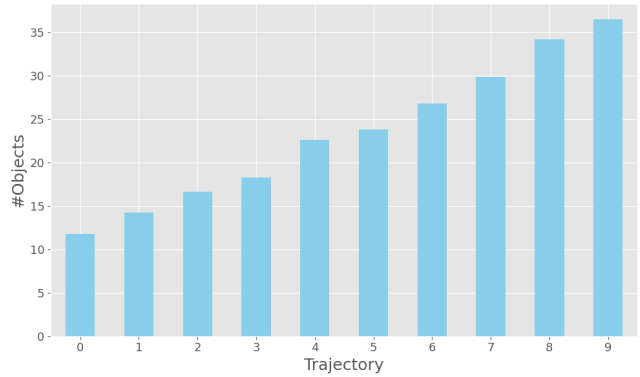
Domain	Operators	Predicates	Types	Const.	Operators arity		Predicates arity		Neg. pre.	Inj. ass.
					min	max	min	max		
barman	12	15	9	no	2	6	1	2	no	no
blocksworld	4	5	1	no	1	2	0	2	no	no
childsnack	6	13	6	yes	2	4	1	2	no	no
depots	5	6	9	no	3	4	1	2	no	no
elevators	6	8	5	no	3	5	2	2	no	no
ferry	3	5	2	no	2	2	0	2	no	no
floortile	7	10	3	no	3	4	1	2	no	no
goldminer	7	12	1	no	1	2	0	2	no	no
grippers	3	4	4	no	3	4	2	3	no	no
matchingbw	10	10	2	no	2	3	1	2	no	no
miconic	4	6	2	no	2	2	1	2	no	no
nomystery	3	6	5	no	3	6	2	3	no	no
npuzzle	1	3	2	no	3	3	1	2	no	no
parking	4	5	2	no	3	3	1	2	no	no
rovers	9	25	7	no	2	6	1	3	no	no
satellite	5	8	4	no	2	4	1	2	no	no
sokoban	2	4	3	no	3	5	1	3	no	no
spanner	3	6	5	no	3	4	1	2	no	no
tpp	4	7	7	no	3	7	2	3	no	no
transport	3	5	6	no	3	5	2	2	no	no

Table 3: Details about the domains available in the proposed benchmark.

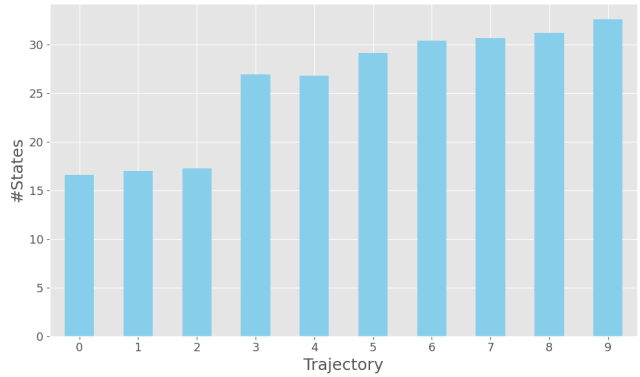
8.2 Training Set T_{train} Statistics

Domain	# States	# Objects	# Lifted Actions
barman	35.80	25.10	34.80
blocksworld	23.00	7.50	22.00
childsnack	25.50	28.40	24.50
depots	21.60	28.30	20.60
elevators	25.80	14.50	24.80
ferry	27.60	14.70	26.60
floortile	35.90	53.70	34.90
goldminer	30.00	21.00	29.00
grippers	15.50	17.60	14.50
matchingbw	25.00	9.50	24.00
miconic	21.00	10.80	20.00
nomystery	19.80	16.80	18.80
npuzzle	30.00	29.80	29.00
parking	21.00	12.10	20.00
rovers	30.00	25.10	29.00
satellite	24.50	25.10	23.50
sokoban	26.40	60.80	25.40
spanner	20.30	21.30	19.30
tpp	30.00	17.30	29.00
transport	28.20	29.60	27.20

Table 4: Average number of states, objects, and lifted actions among the set of 10 trajectories for every domain.



(a) Number of objects in every trajectory of T_{train} averaged among all domains.

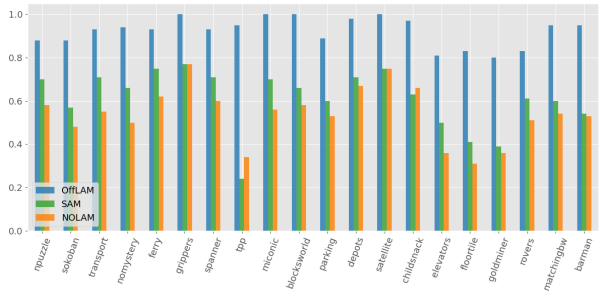


(b) Number of states in every trajectory of T_{train} averaged among all domains.

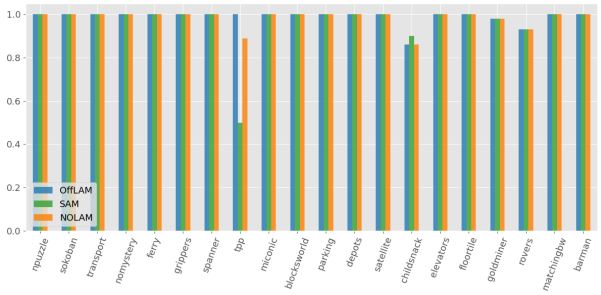
8.3 Detailed Experimental Results

Domain	Syntactic		Applicability		Predicted effects		Solvability % ↑	False plans % ↓
	P↑	R↑	P↑	R↑	P↑	R↑		
barman	0.95 [2]	1.0 [1,2,3]	1.0 [1,2,3]	0.94 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
blocksworld	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
childsnack	0.97 [2]	0.9 [1]	1.0 [1]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [1]	0 [1]
depots	0.98 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
elevators	0.81 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
ferry	0.93 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
floortile	0.83 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
goldminer	0.8 [2]	0.98 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [1,3]	0 [1,3]
grippers	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
matchingbw	0.95 [2]	1.0 [1,2,3]	1.0 [1,2,3]	0.98 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
miconic	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
nomystery	0.94 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
npuzzle	0.88 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
parking	0.89 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
rovers	0.83 [2]	0.93 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
satellite	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
sokoban	0.88 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
spanner	0.93 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
tpp	0.95 [2]	1.0 [2]	1.0 [1,2,3]	1.0 [2]	1 [1,2,3]	1 [1,2,3]	1.0 [2]	0 [1,2,3]
transport	0.93 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1 [1,2,3]	1 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]

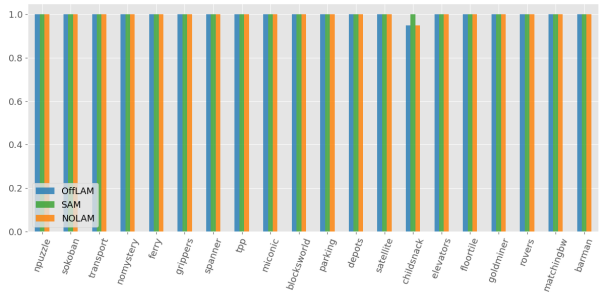
Table 5: Best metric values for every domain obtained among SAM, OffLAM, NOLAM, respectively denoted by 1,2, and 3; P indicates the precision and R the recall, ↑ (resp. ↓) denotes the higher (resp. lower) the better. The training set T_{train} includes 10 trajectories for every domain.



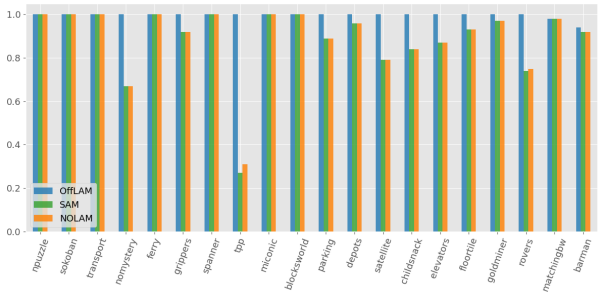
(a) Syntactic precision



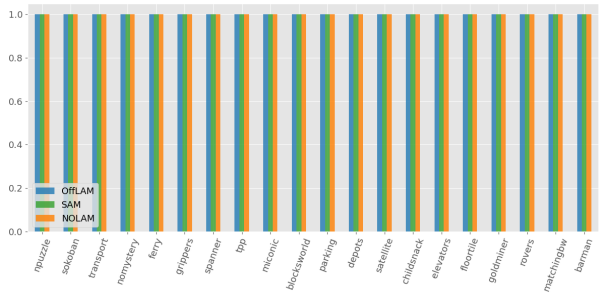
(b) Syntactic recall



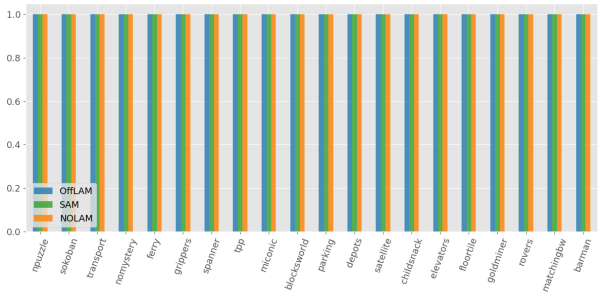
(c) Applicability precision



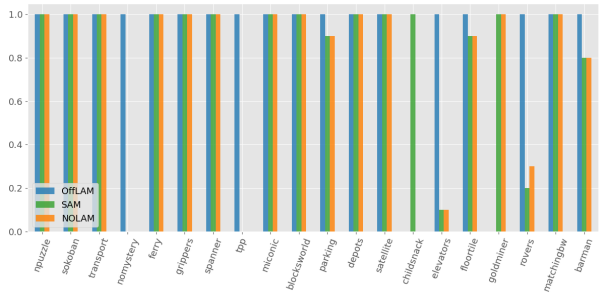
(d) Applicability recall



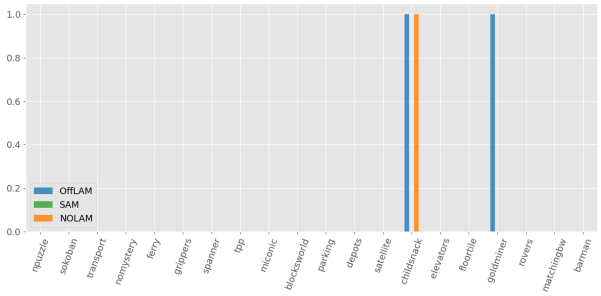
(e) Predicted effects precision



(f) Predicted effects recall



(g) Problem solving ratio



(h) False plans ratio

Figure 3: Evaluation metric values for the models learned by SAM, OffLAM, NOLAM in every benchmark domain. The training set T_{train} includes 10 trajectories for every domain.

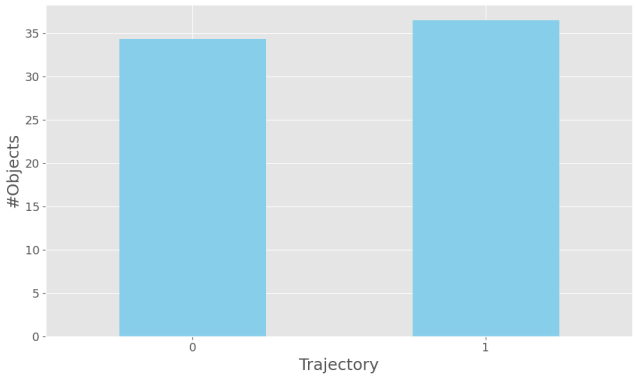
8.4 Additional Experimental Results

We also conducted an additional experiments where we generated a training set $\hat{T}_{train}$ with only 2 trajectories with 10 states for every domain. To generate this dataset, we produced heuristic plan by considering the 2 most complex problem settings (in terms of number of objects) adopted for generating the training set with 10 trajectories for every domain.

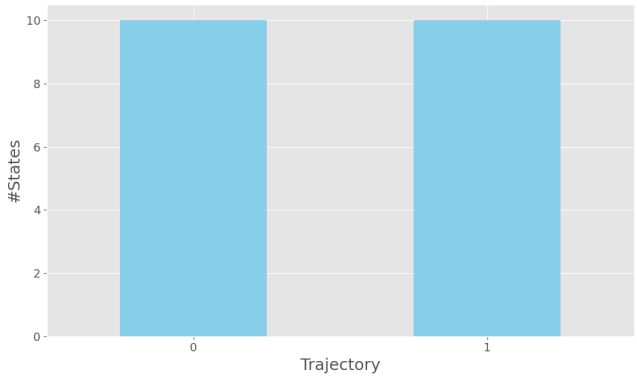
8.5 Training Set $\hat{T}_{train}$ Statistics

Domain	# Objects	# States	# Lifted Actions
barman	33.50	9.00	10.00
blocksworld	11.50	9.00	10.00
childsnack	38.50	9.00	10.00
depots	40.50	9.00	10.00
elevators	17.50	9.00	10.00
ferry	23.00	9.00	10.00
floortile	73.00	9.00	10.00
goldminer	33.00	9.00	10.00
grippers	27.50	9.00	10.00
matchingbw	13.50	9.00	10.00
miconic	15.00	9.00	10.00
nomystery	23.50	9.00	10.00
npuzzle	49.00	9.00	10.00
parking	18.00	9.00	10.00
rovers	38.50	9.00	10.00
satellite	39.50	9.00	10.00
sokoban	107.00	9.00	10.00
spanner	32.50	9.00	10.00
tpp	24.50	9.00	10.00
transport	48.50	9.00	10.00

Table 6: Average number of states, objects, and lifted actions among the set of 2 trajectories for every domain.



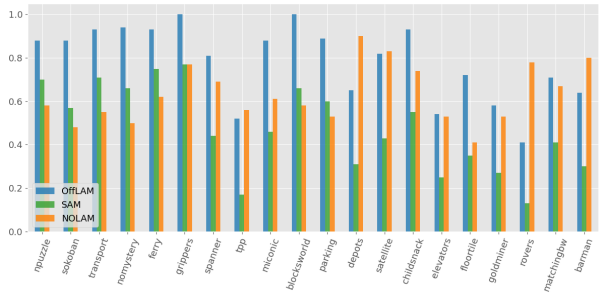
(a) Number of objects in every trajectory of $\hat{T}_{train}$ averaged among all domains.



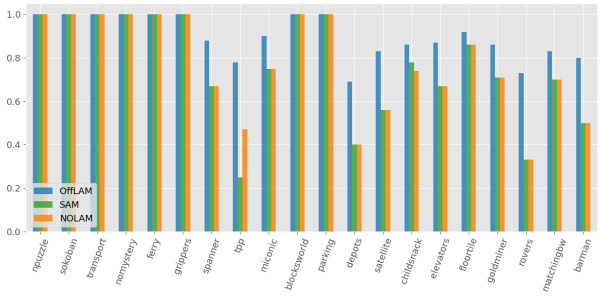
(b) Number of states in every trajectory of $\hat{T}_{train}$ averaged among all domains.

Domain	Syntactic		Applicability		Predicted effects		Solvability % ↑	False plans % ↓
	P↑	R↑	P↑	R↑	P↑	R↑		
barman	0.8 [3]	0.8 [2]	1.0 [1,2]	0.98 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
blocksworld	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
childsnack	0.93 [2]	0.86 [2]	1.0 [1]	0.91 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
depots	0.9 [3]	0.69 [2]	1.0 [1,2]	0.96 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
elevators	0.54 [2]	0.87 [2]	1.0 [1,2]	0.62 [3]	1.0 [1,2,3]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
ferry	0.93 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
floortile	0.72 [2]	0.92 [2]	1.0 [1,2]	0.93 [3]	1.0 [1,2]	1.0 [1,2]	0.5 [1]	0 [1,2,3]
goldminer	0.58 [2]	0.86 [2]	1.0 [1,2]	0.79 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
grippers	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
matchingbw	0.71 [2]	0.83 [2]	1.0 [1,2]	0.92 [3]	1.0 [1,2]	1.0 [1,2]	0.3 [1]	0 [1,2,3]
miconic	0.88 [2]	0.9 [2]	1.0 [1,2]	0.84 [3]	1.0 [1,2,3]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
nomystery	0.94 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	0 [1,2,3]
npuzzle	0.88 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
parking	0.89 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [2]	0 [1,2,3]
rovers	0.78 [3]	0.73 [2]	1.0 [1,2]	0.74 [3]	1.0 [1,2,3]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
satellite	0.83 [3]	0.83 [2]	1.0 [1,2]	0.87 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
sokoban	0.88 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]
spanner	0.81 [2]	0.88 [2]	1.0 [1,2]	1.0 [3]	1.0 [1,2]	1.0 [1,2]	0.0 [1,2,3]	0 [1,2,3]
tpp	0.56 [3]	0.78 [2]	1.0 [1,2]	0.75 [3]	1.0 [1,2]	1.0 [1]	0.0 [1,2,3]	0 [1,2,3]
transport	0.93 [2]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	1.0 [1,2,3]	0 [1,2,3]

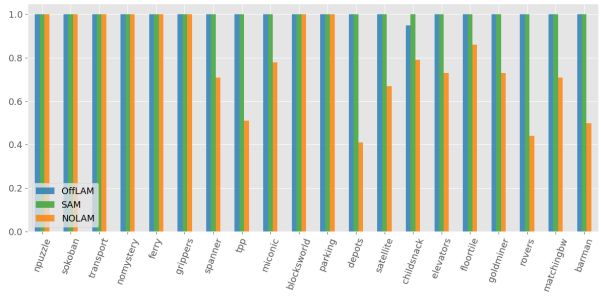
Table 7: Best metric values for every domain obtained among SAM, OffLAM, NOLAM, respectively denoted by 1,2, and 3; P indicates the precision and R the recall, ↑ (resp. ↓) denotes the higher (resp. lower) the better. The training set $\hat{T}_{train}$ includes 2 trajectories for every domain.



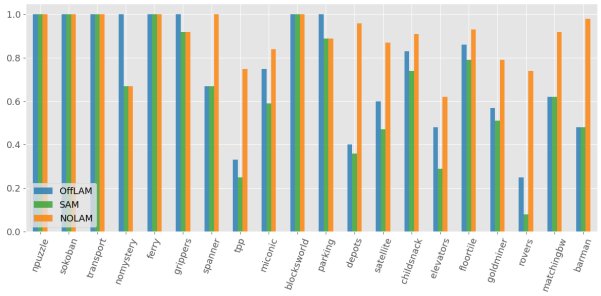
(a) Syntactic precision



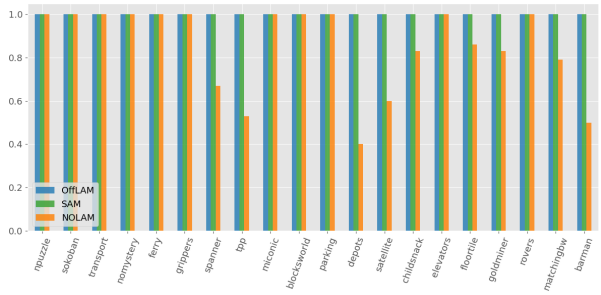
(b) Syntactic recall



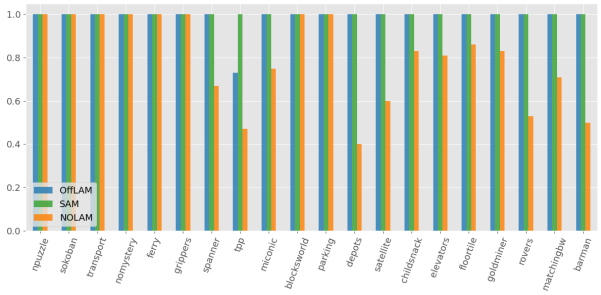
(c) Applicability precision



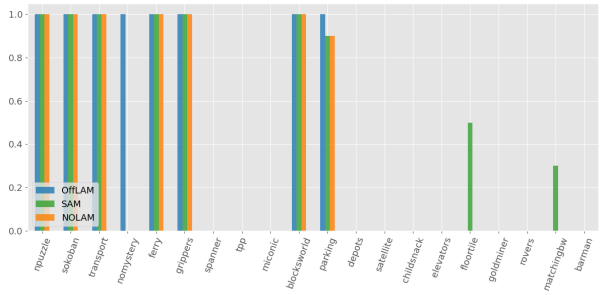
(d) Applicability recall



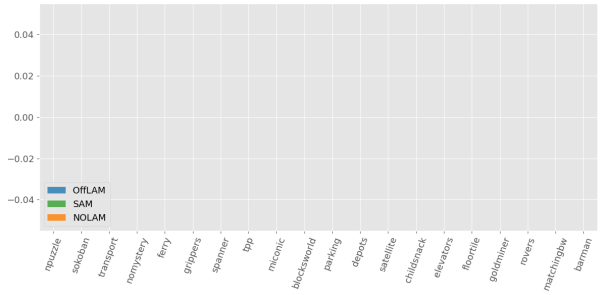
(e) Predicted effects precision



(f) Predicted effects recall



(g) Problem solving ratio



(h) False plans ratio

Figure 5: Evaluation metrics when learning from a training set $\hat{T}_{train}$ with 2 traces for every domain.