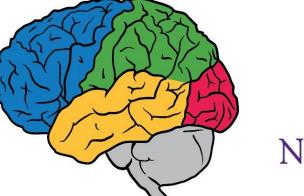
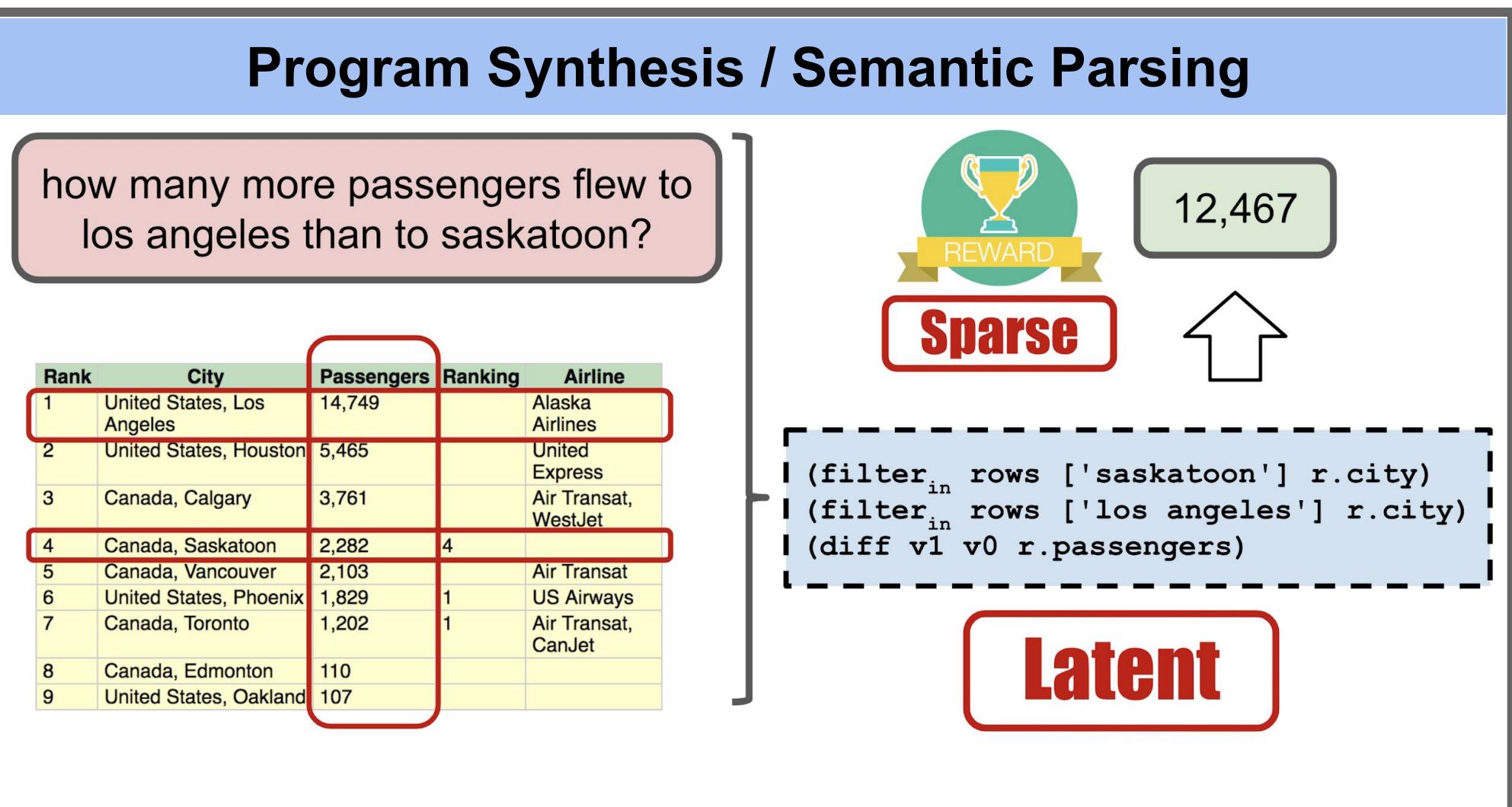
Memory Augmented Policy Optimization (MAPO) for Program Synthesis and Semantic Parsing

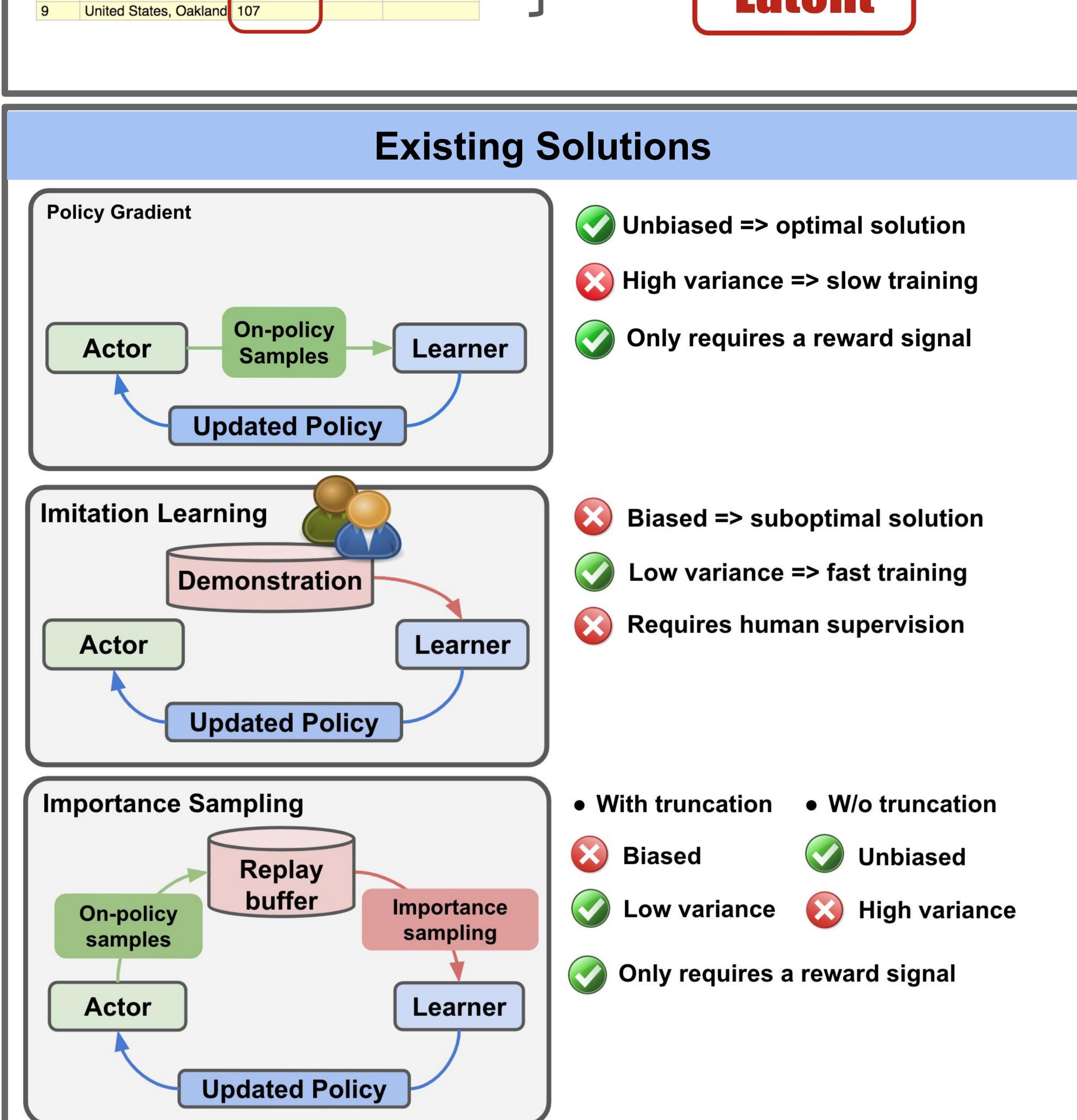






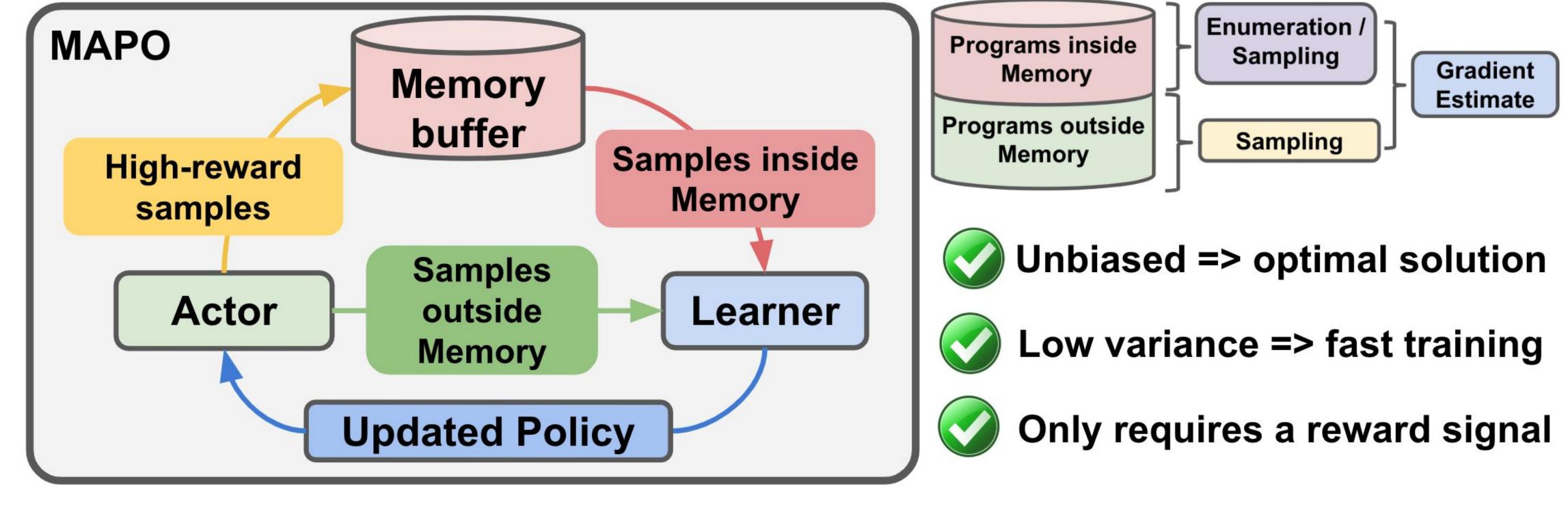
Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc Le, Ni Lao





Memory Augmented Policy Optimization

 MAPO incorporates a memory of promising samples an unbiased gradient estimate with low variance.



 Decompose the expected return objective into weighted sum of two expectations inside and outside the memory.

$$\mathcal{O}_{\mathrm{ER}}(\theta) = \pi_{\mathcal{B}} \underbrace{\mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^{+}(\mathbf{a})} R(\mathbf{a})}_{\text{Expectation inside } \mathcal{B}} + \underbrace{(1 - \pi_{\mathcal{B}}) \underbrace{\mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^{-}(\mathbf{a})} R(\mathbf{a})}_{\text{Expectation outside } \mathcal{B}}$$

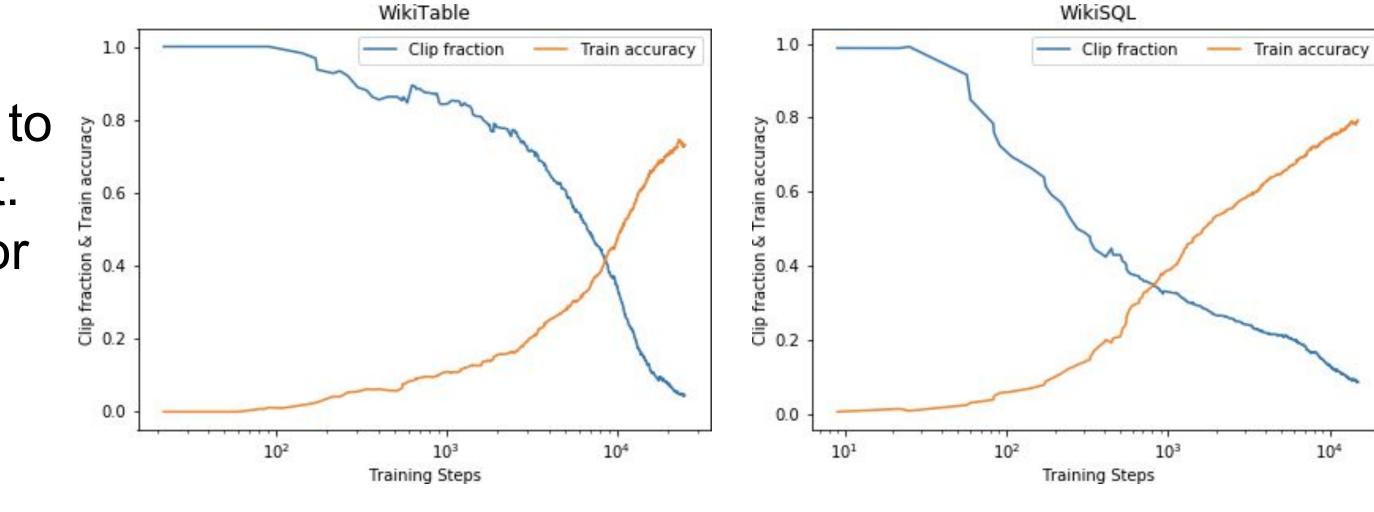
$$\nabla_{\theta} \mathcal{O}_{ER}(\theta) = \pi_{\mathcal{B}} \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^{+}(\mathbf{a})} \nabla \log \pi_{\theta}(\mathbf{a}) R(\mathbf{a}) + (1 - \pi_{\mathcal{B}}) \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^{-}(\mathbf{a})} \nabla \log \pi_{\theta}(\mathbf{a}) R(\mathbf{a})$$

 ${\cal B}$ denotes the memory buffer. $\pi_{ heta}^+$ and $\pi_{ heta}^-$ denotes the renormalized probability.

Memory weight clipping

- Force the training to pay attention to the memory by clipping the weight.
- Trade off bias in the initial stage for faster training.

 $\pi_{\mathcal{B}}^c = \max(\pi_{\mathcal{B}}, \alpha)$

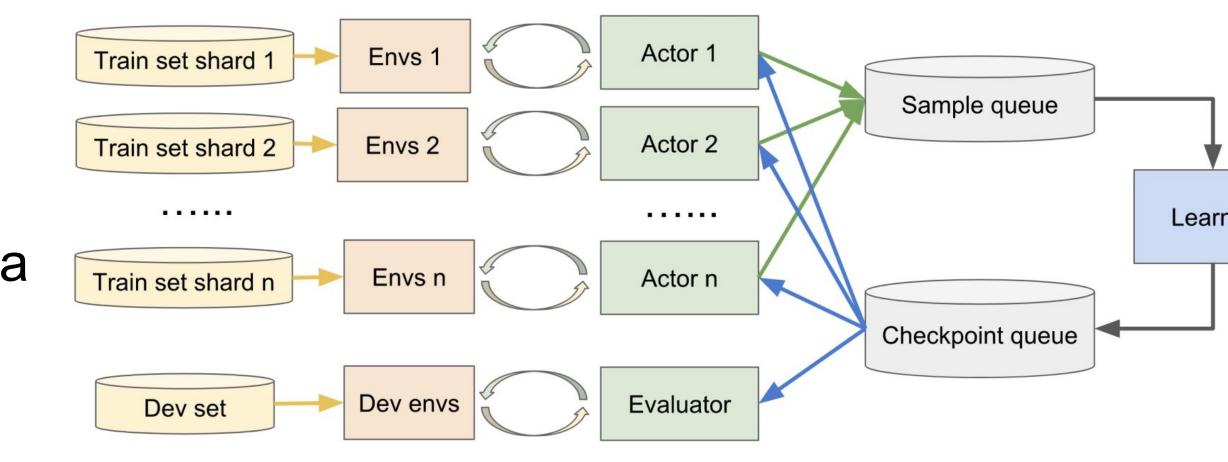


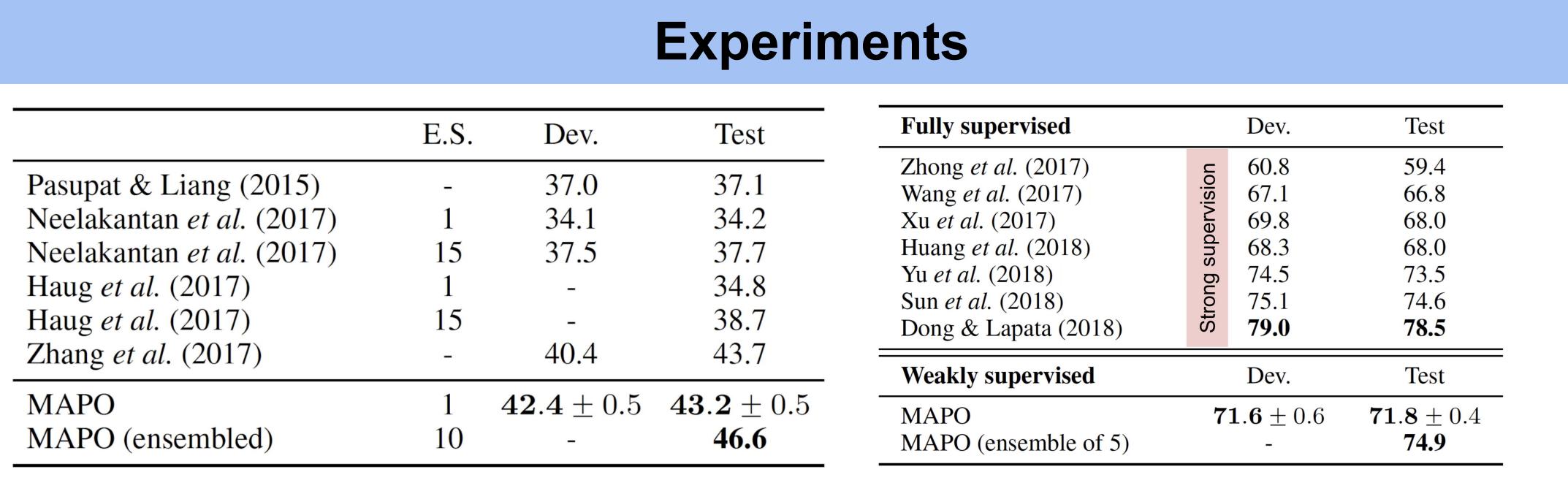
Systematic exploration

- Use a bloom filter to force the exploration to generate new programs.
- Trade-off memory for more efficient exploration.

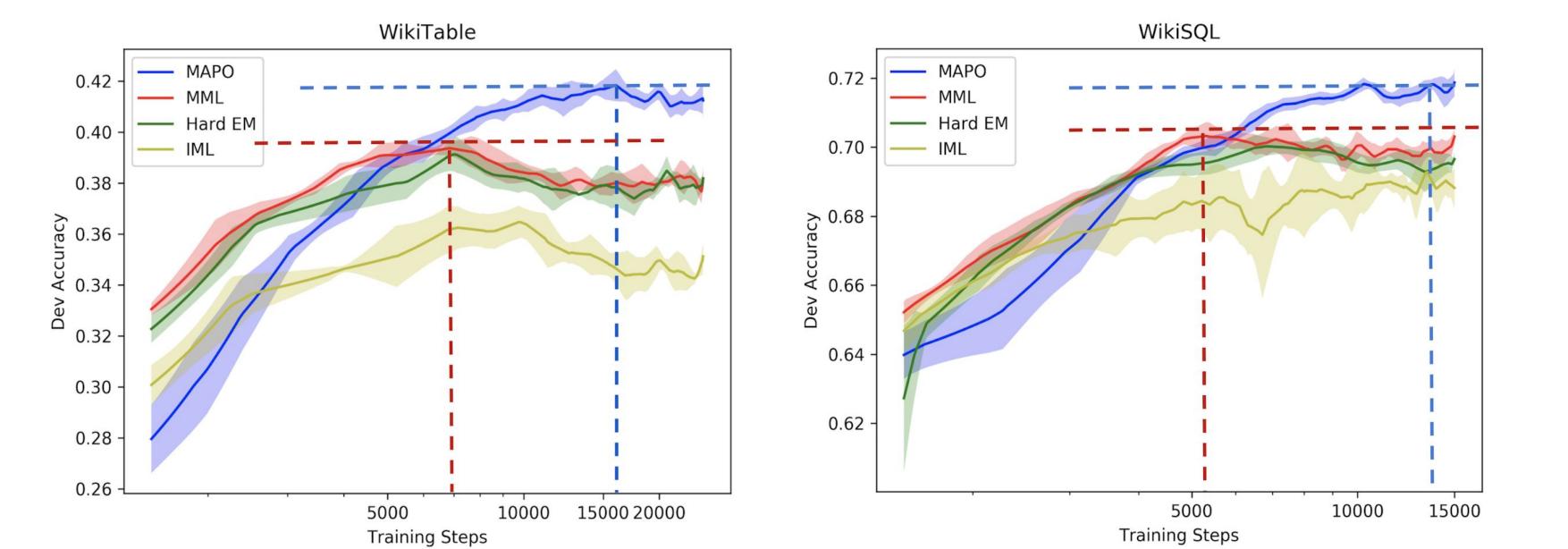
Distributed sampling

- \circ Distribute the cost of computing $\pi_{\mathcal{B}}$ and sampling into the actors.
- Multiple actors each interacting with a shard of training set and send samples to a learner.





- First RL-based state-of-the-art approach on WikiTableQuestions.
- Competitive to state-of-the-art on WikiSQL, which uses strong supervision (the ground truth programs), while MAPO only uses weak supervision (the final answers).



- MAPO converges slower than maximum likelihood training, but reaches a better solution.
- REINFORCE doesn't make much progress (<10% accuracy).
- Spurious programs: right answer for the wrong reason

Rank	Nation	Gold	Silver	Bronze	Total
1	Nigeria	14	12	9	35
2	Algeria	9	4	4	17
3	Kenya	8	11	4	23
4	Ethiopia	2	4	7	13
5	Ghana	2	2	2	6
6	Ivory Coast	2	1	3	6
7	Egypt	2	1	0	3
8	Senegal	1	1	5	7

Which nation won the most silver medal?

Correct program:

(argmax rows "Silver") (hop v1 "Nation")

• Spurious programs:

(argmax rows "Gold") (argmax rows "Bronze") (hop v1 "Nation") (hop v1 "Nation")

Comparison of MAPO, MML, IML with a simplified example

	Ques	stion 1	Question 2		
	correct	spurious	spurious	spurious	
Iterative Maximum Likelihood (IML)	0.5	0.5	0.5	0.5	
Maximum Marginal Likelihood (MML)	0.8	0.2	0.5	0.5	
MAPO	0.6	0.15	0.1	0.1	
Model Probability	0.6	0.15	0.1	0.1	