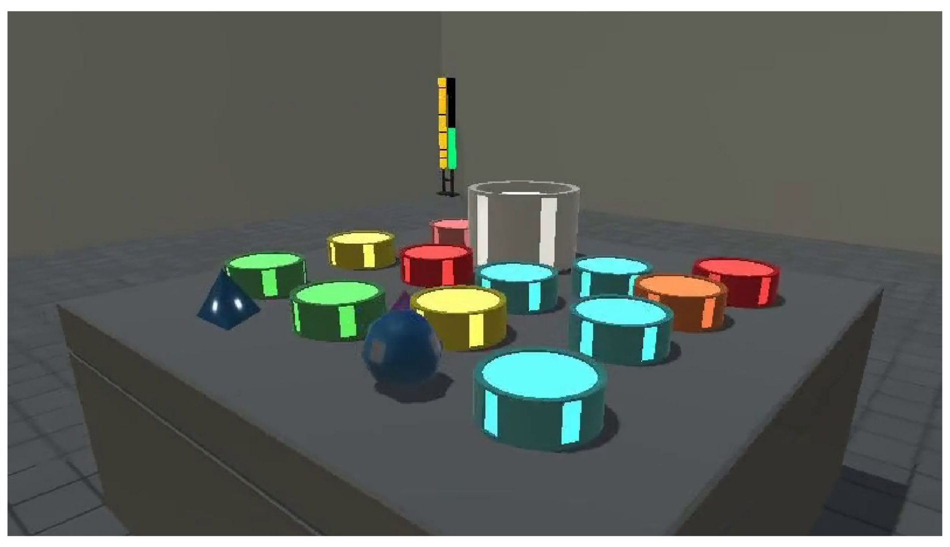
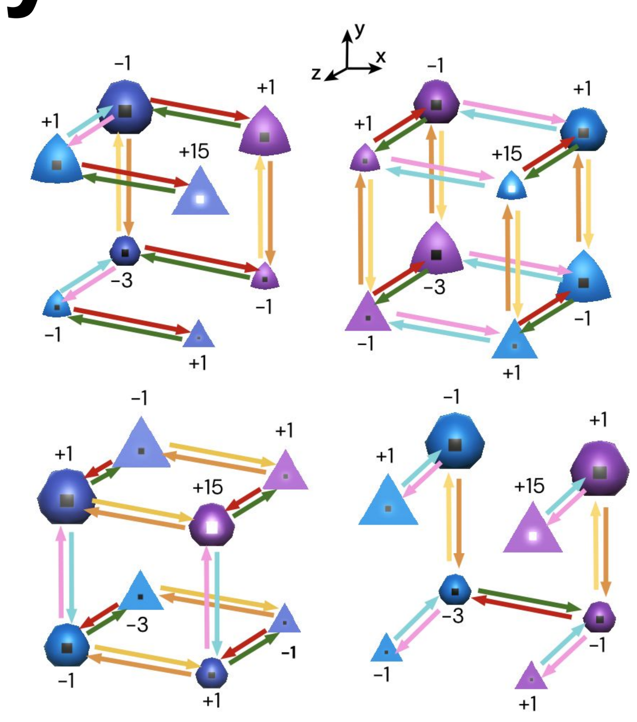
Learning How to Infer Partial MDPs for In-Context Adaptation and Exploration

Chentian Jiang, Rosemary Ke, Hado van Hasselt

How can an RL agent generalize across tasks?
How can it explore and adapt *in-context* (i.e., without gradient updates) in new tasks?

Meta-RL Benchmark: Symbolic Alchemy (Wang et al., 2021)

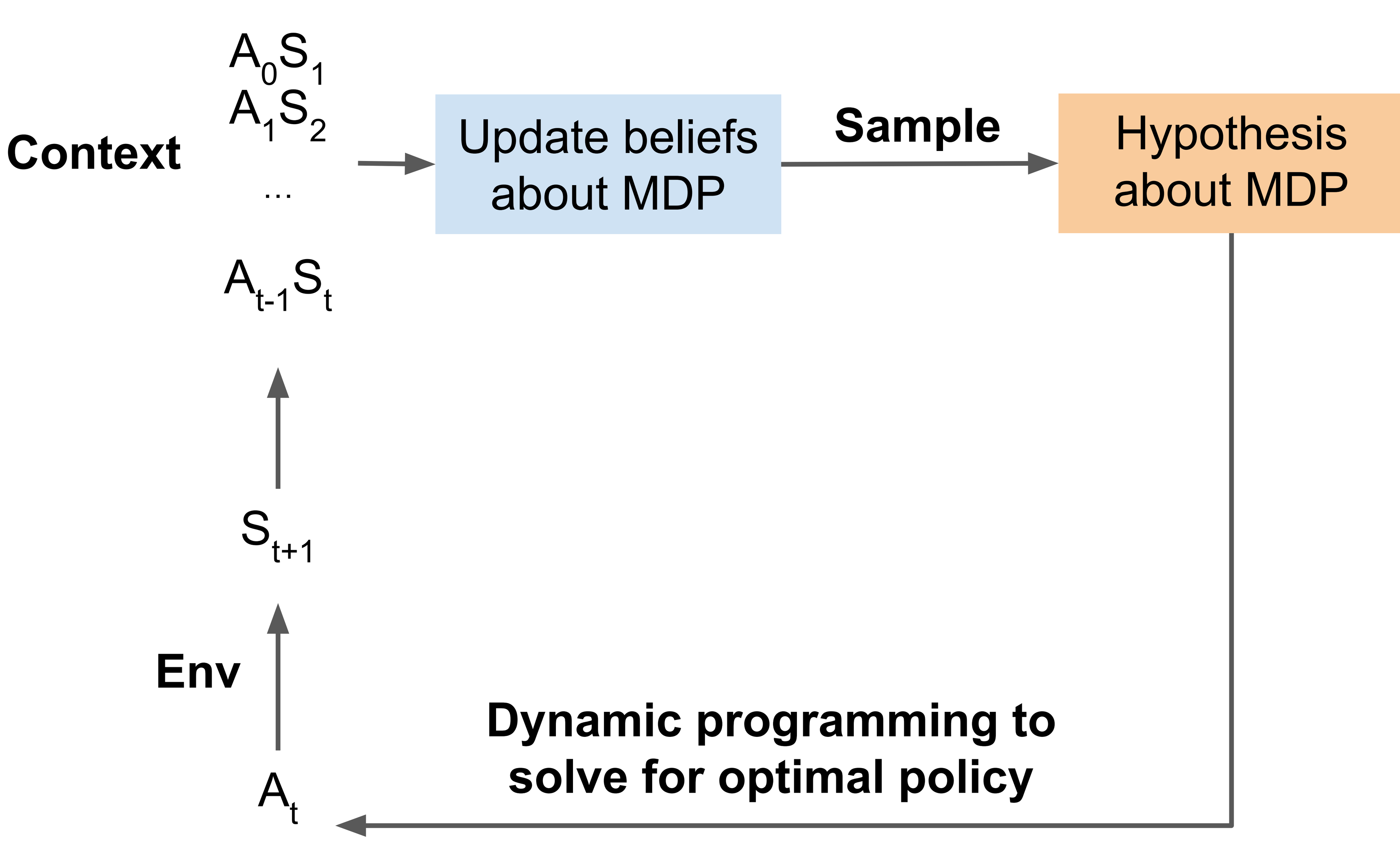




Problem (in our modified version):
- Each task can be represented as a partial MDP graph (see cubes to the left).
- Graph edges change across episodes, i.e., the transition function changes.
- Only 200 time steps to explore and adapt to a new episode.

Posterior Sampling Framework

(Thompson, 1933; Strens, 2000; Osband et al., 2013)



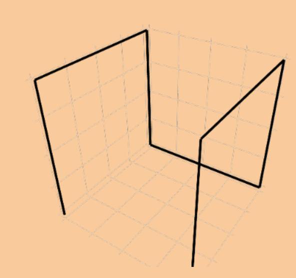
The original posterior sampling procedure is limited because:

Sample can be a large MDP
➜ Expensive dynamic prog.

Update via Bayesian inference
- Expensive
- Unknown prior

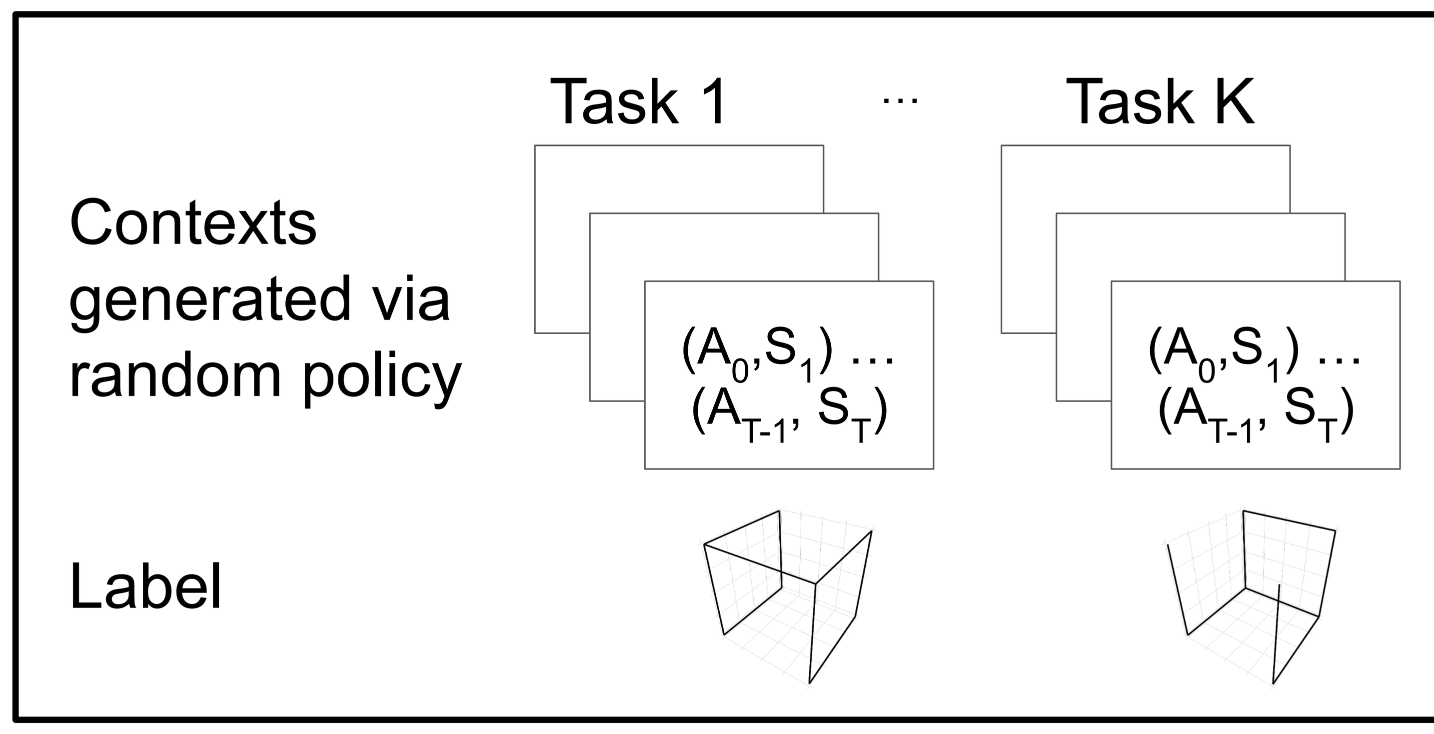
Our Contributions

Sample is a **partial** MDP graph
➜ Cheap dynamic prog.

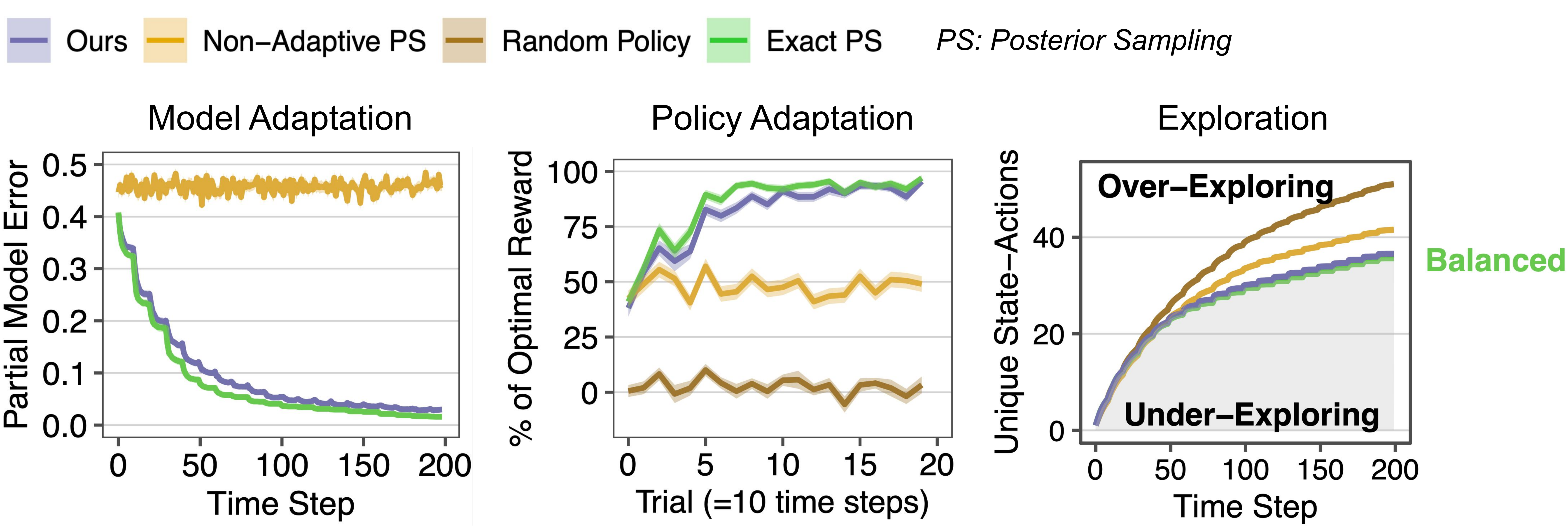


Posterior update *p(graph edges | context)* is **approximated via a transformer and training tasks**
- Replace exact Bayesian inference

Offline Transformer Training (Ke et al., 2022)



Results in Held Out Tasks (no gradient updates)



We almost match the adaptation speed, model accuracy, rewards, and exploration behavior of an exact Posterior Sampling oracle.
Future: How can we learn partial MDP graphs for other environments?