EMNLP 2022

UNIVERSITY OF MICHIGAN

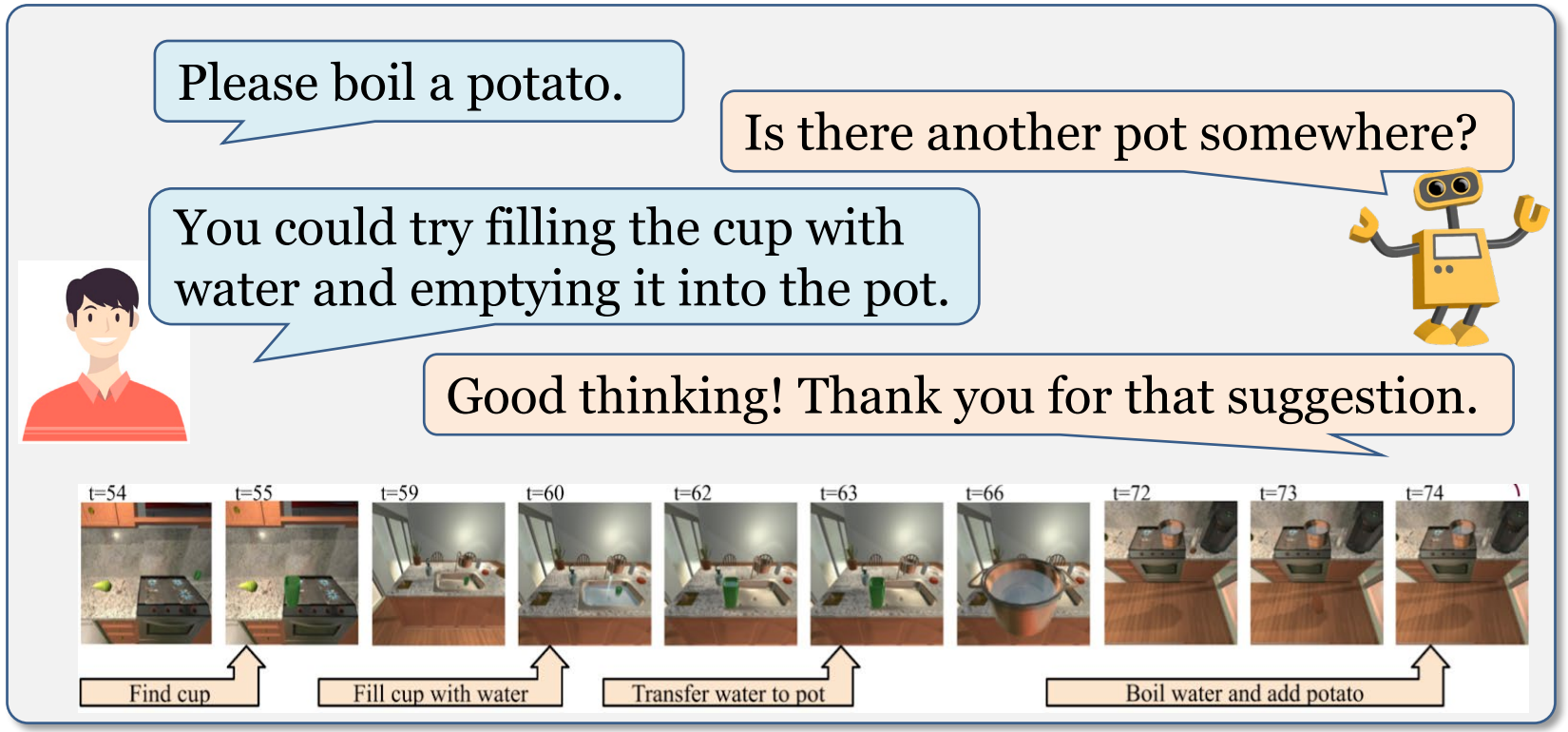
DANLI: Deliberative Agent for Following Natural Language Instructions

Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj
Ziqiao Ma, Keunwoo Peter Yu, Yuwei Bao, and Joyce Chai
Situated Language and Embodied Dialog (SLED) Lab
University of Michigan

SLED

Problem Definition

- TEACh[1]: The agent is given a dialog history as input, and is expected to execute a sequence of actions to achieve the goal set out by the human commander.

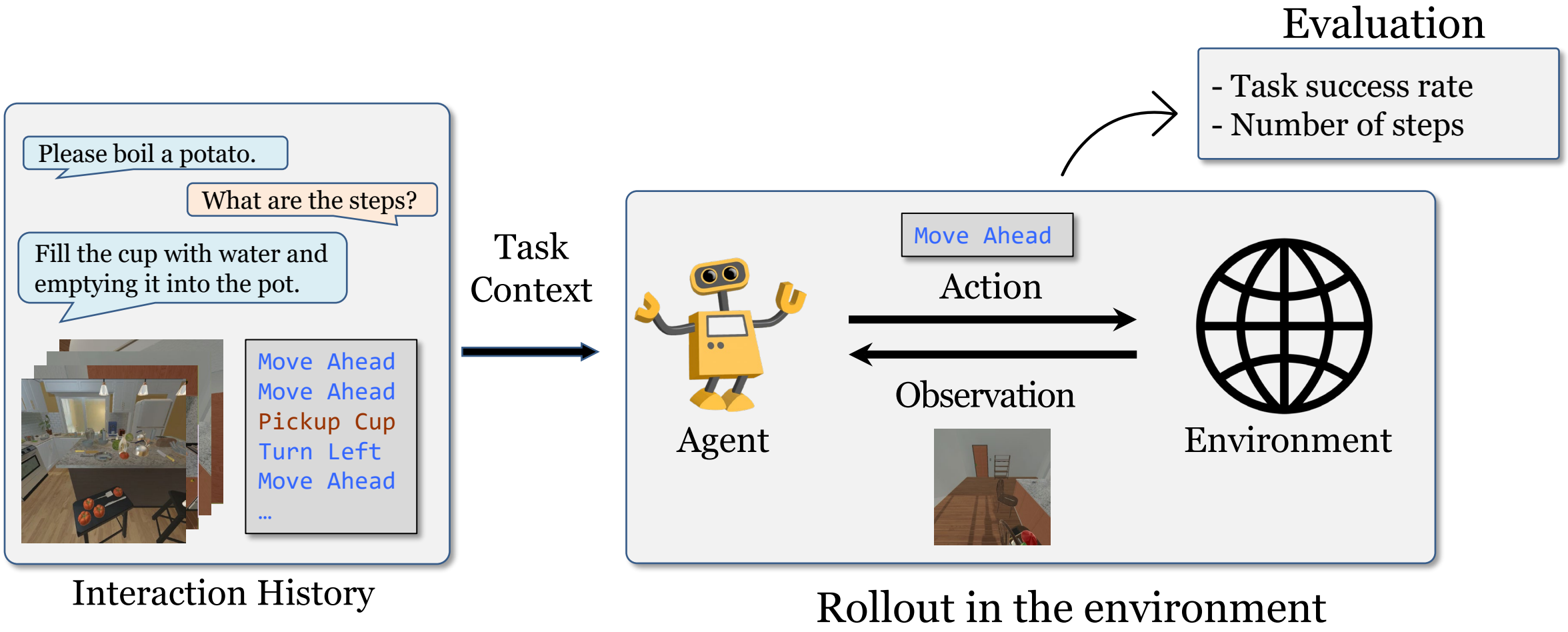


TEACh Tasks

- Water Plant
- Make Coffee
- Clean All X
- Put All X On Y
- Boil Potato
- Make Plate of Toast
- N Slices Of X In Y
- Put All X In One Y
- N Cooked X Slices In Y
- Prepare Sandwich
- Prepare Salad
- Prepare Breakfast

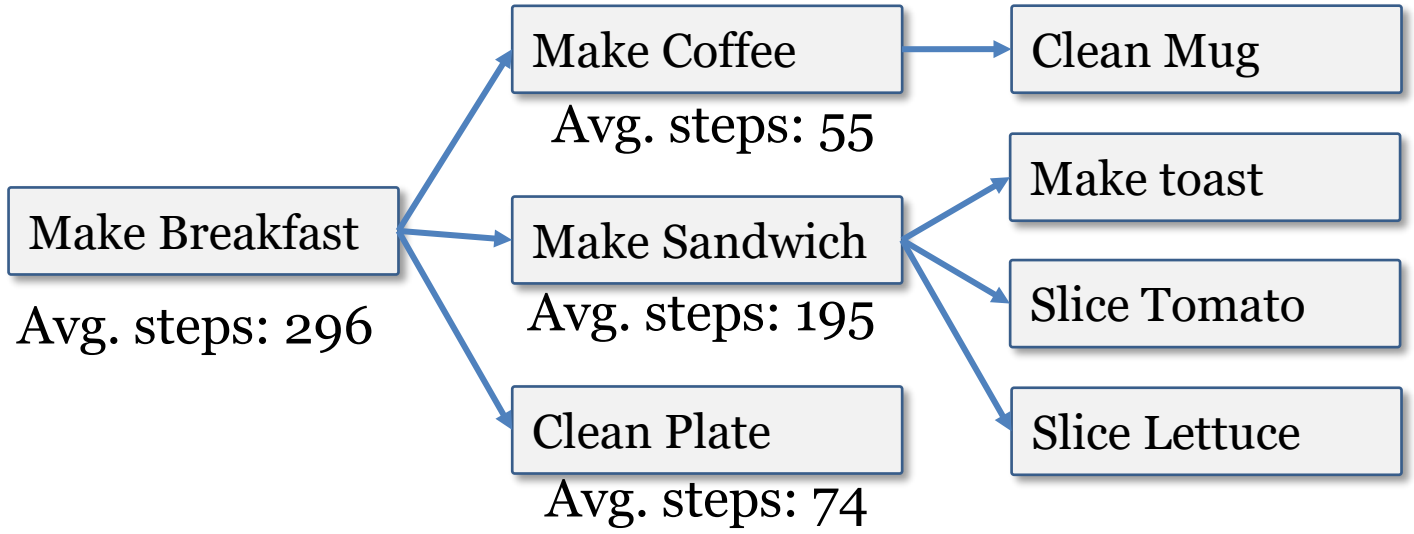
[1] Padmakumar A, Thomason J, Shrivastava A, Lange P, Narayan-Chen A, Gella S, Piramuthu R, Tur G, Hakkani-Tur D. Teach: Task-driven embodied agents that chat. AAAI 2022.

Execution From Dialog History (EDH)



Challenges

- Multi-modal situation and task understanding
- Partially observable and high-dimensional state
- Long-horizon and compositional tasks



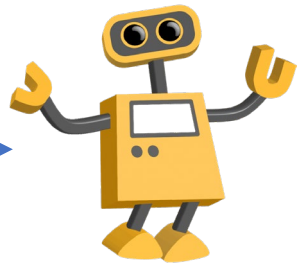
Motivation: From Reactive to Deliberative

Pashevich et al. 2021
Bulks et al. 2021
Zhang and Chai. 2021
Zhou et al. 2021
Singh et al. 2020
...

~ 7% success rate on TEACh-EDH 😢

Reactive

Situation

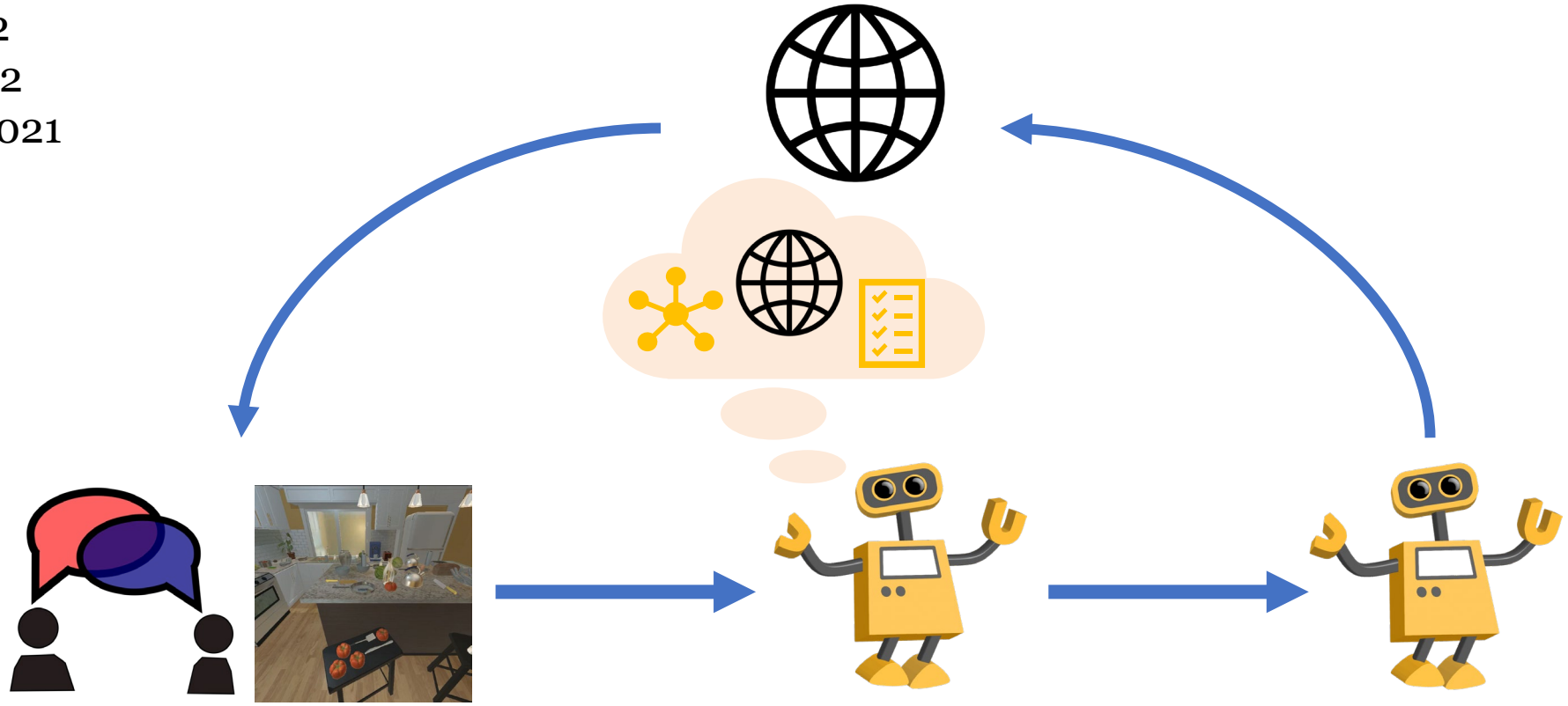


Action

Motivation: From Reactive to Deliberative

Agia et al., 2022
Wang et al., 2022
Srivastava et al., 2021
She et al., 2014
…

Deliberative



Situation

State Tracking & Planning

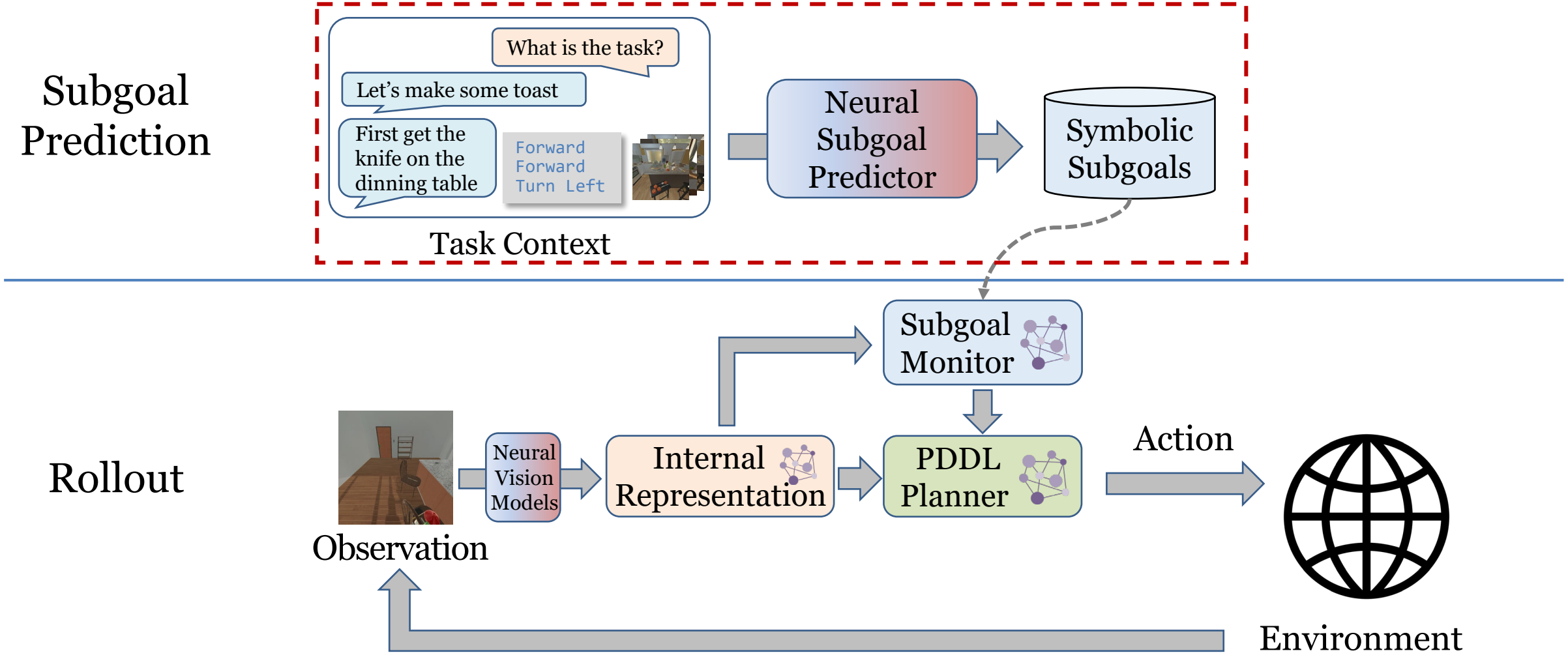
Action

Contributions

Deliberative Agent for following Natural Language Instructions (**DANLI**):

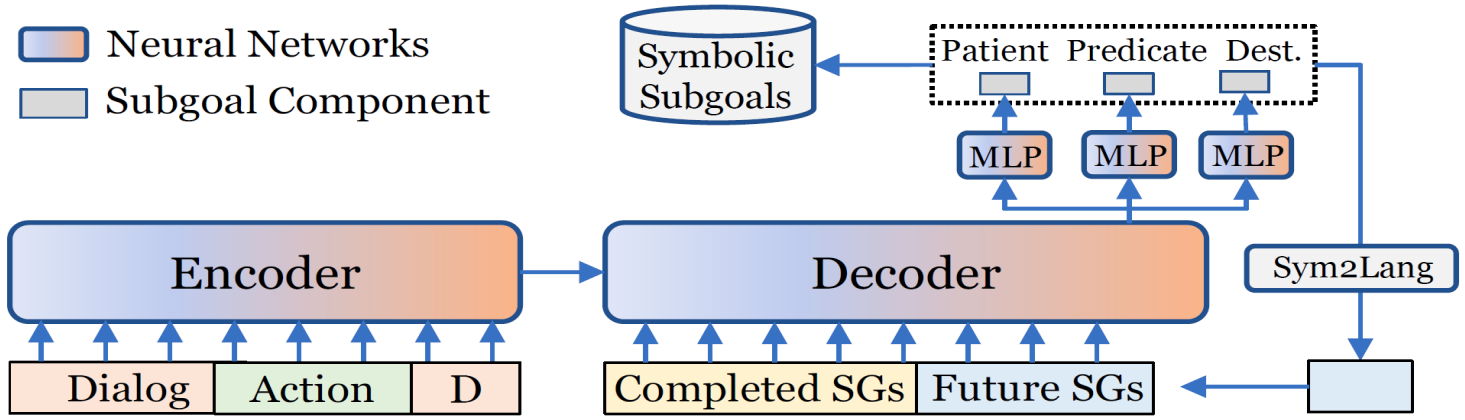
1. **Neural subgoal predictor** uses a language model to predict symbolic subgoals from situated dialog history
2. **Spatial-Symbolic world representation** tracks and grounds objects in a 3D voxel map to facilitate both path and task planning
3. **Online symbolic planner** generates efficient, interpretable plans while allowing online exception handling

Neuro-Symbolic System Architecture



Subgoal Learning From Dialog

- Neural encoder-decoder network for subgoal prediction from interaction history
- Joint progress estimation and future subgoal prediction



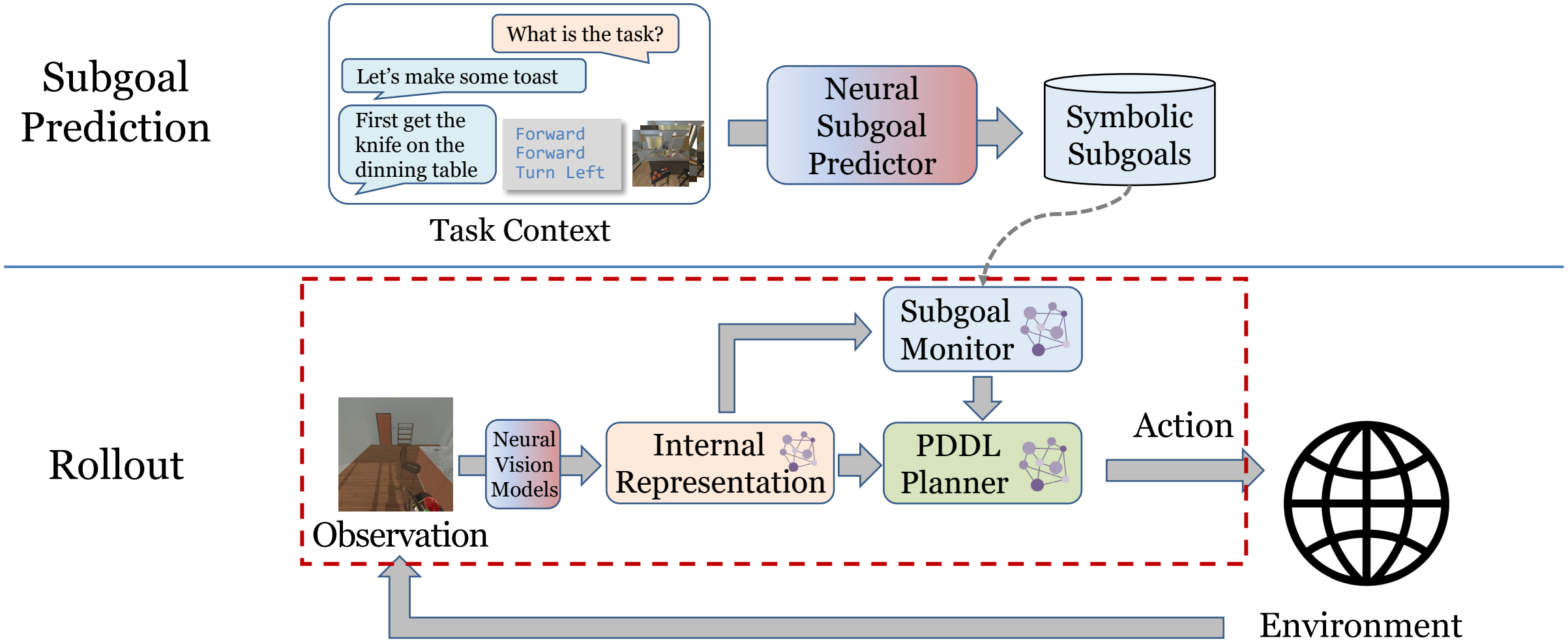
Example Dialog & Action History Input

Follower: Hi. What can I do for you? Commander: Find a cup. Follower go for cup, pick up cup. Commander : Put it on the table.

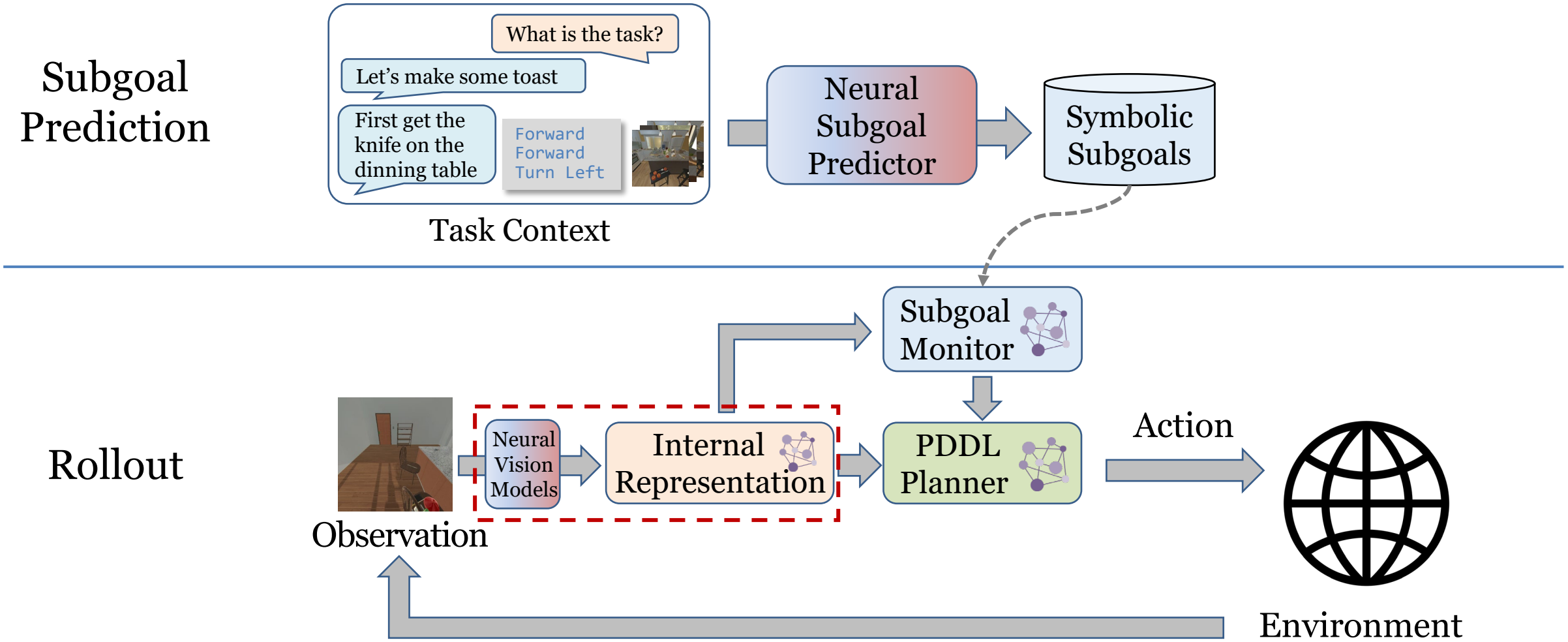
Example Completed & Future Subgoal Output

Completed SG: `(Cup,isPickedUp)`
Future SG: `(Cup,isPlacedTo,Table)`

Neuro-Symbolic System Architecture



Neuro-Symbolic System Architecture

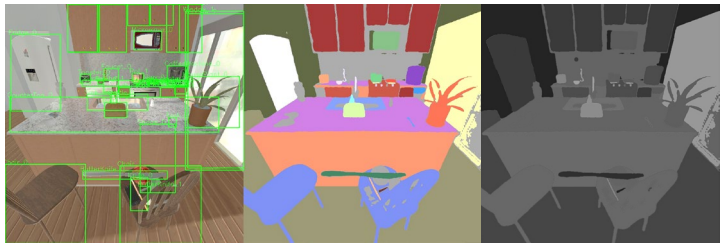


Spatial-Symbolic World Representation

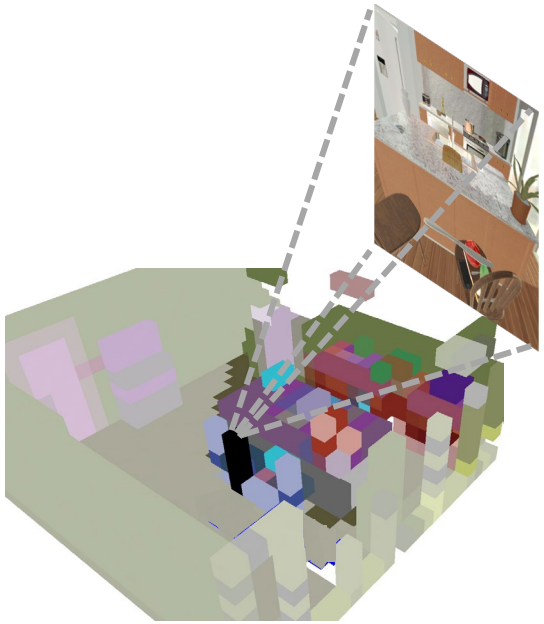
- Neural-powered 3D voxel map construction from ego-centric observations

RGB Observation

Neural Vision Models



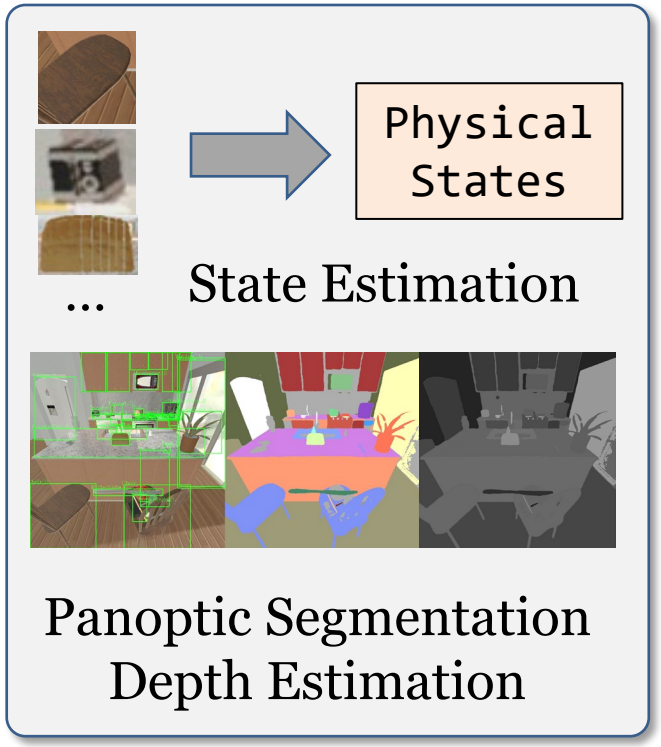
Panoptic Segmentation
Depth Estimation

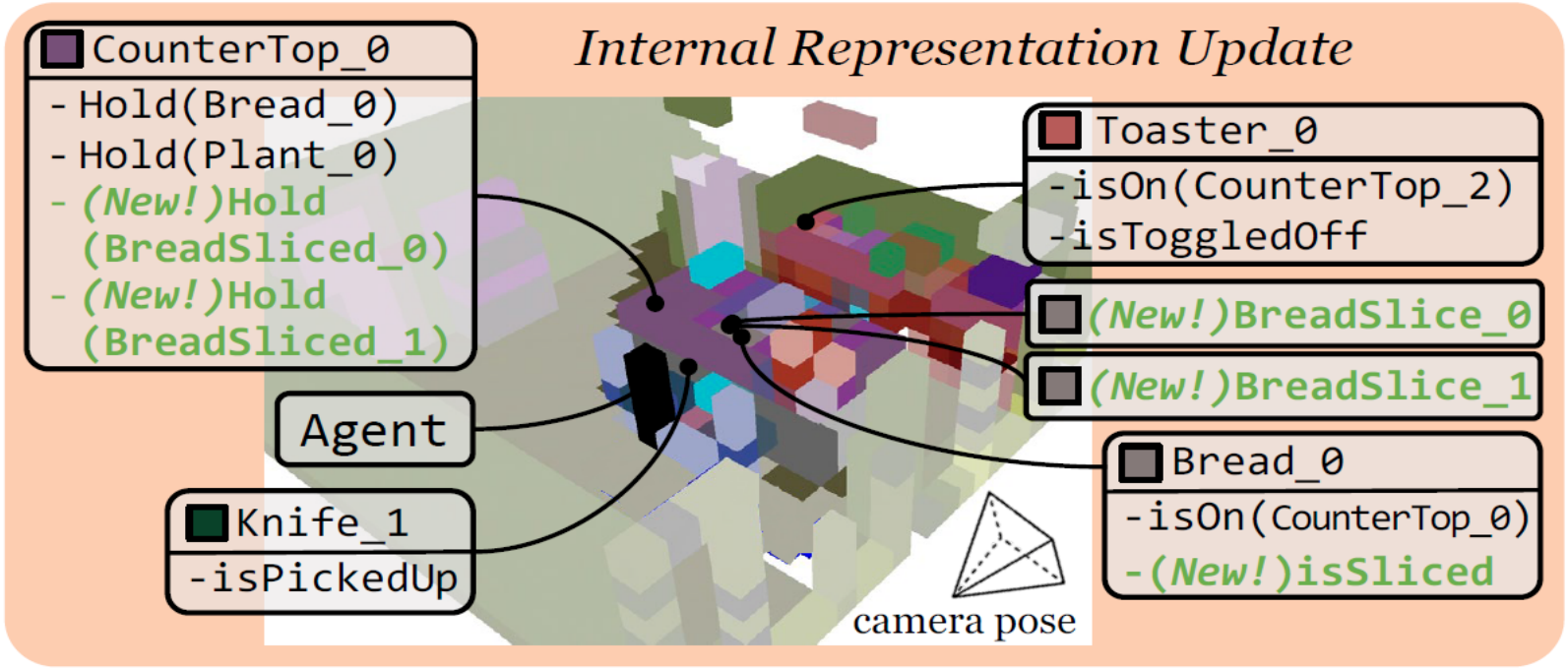


Voxel Map Construction

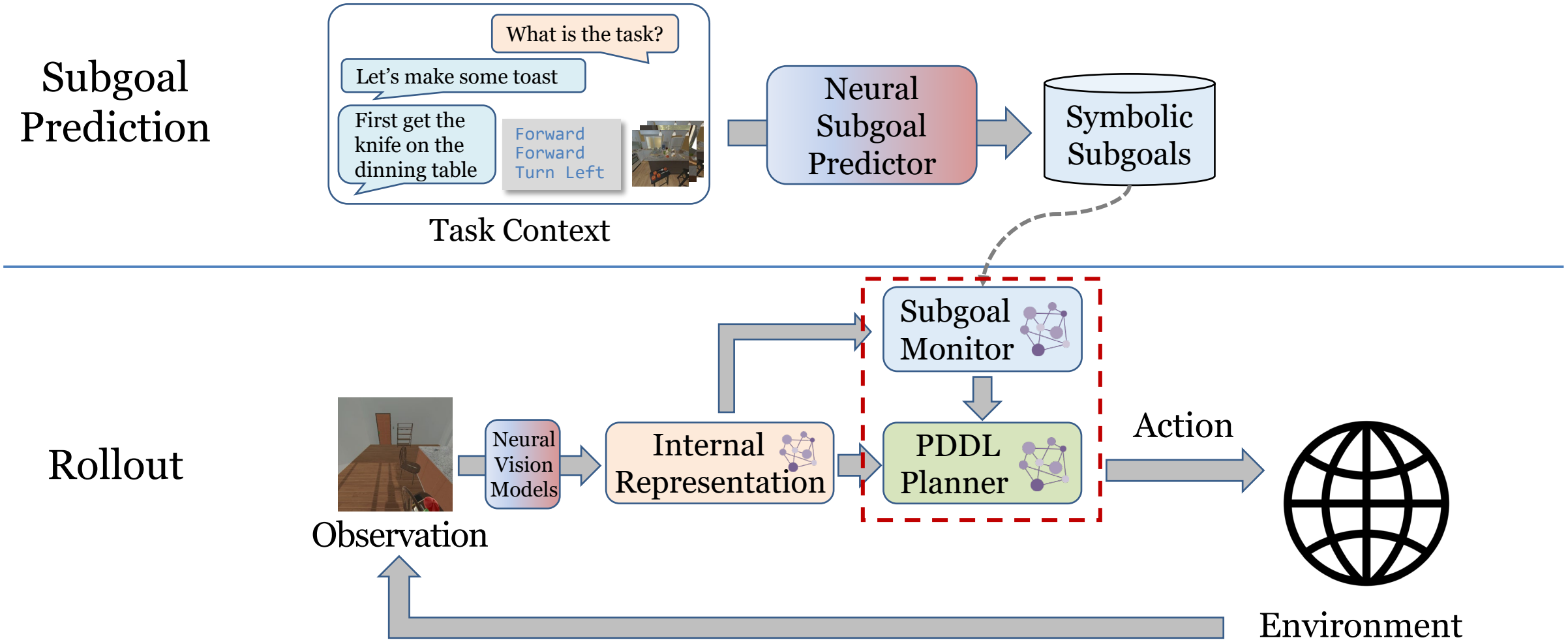
Spatial-Symbolic World Representation

- Recognize physical states of each object instance
- Update the spatial-symbolic world representation at every step

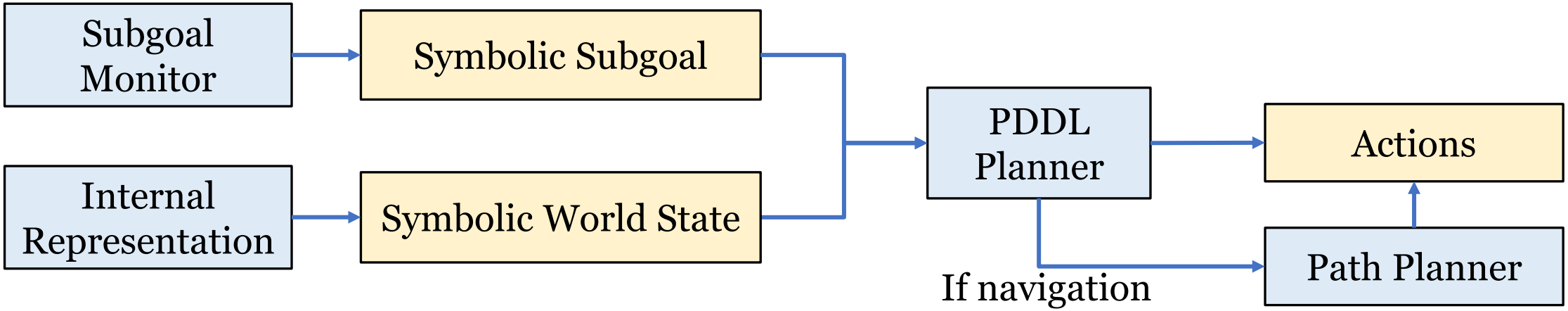




Neuro-Symbolic System Architecture



Symbolic Planning Pipeline



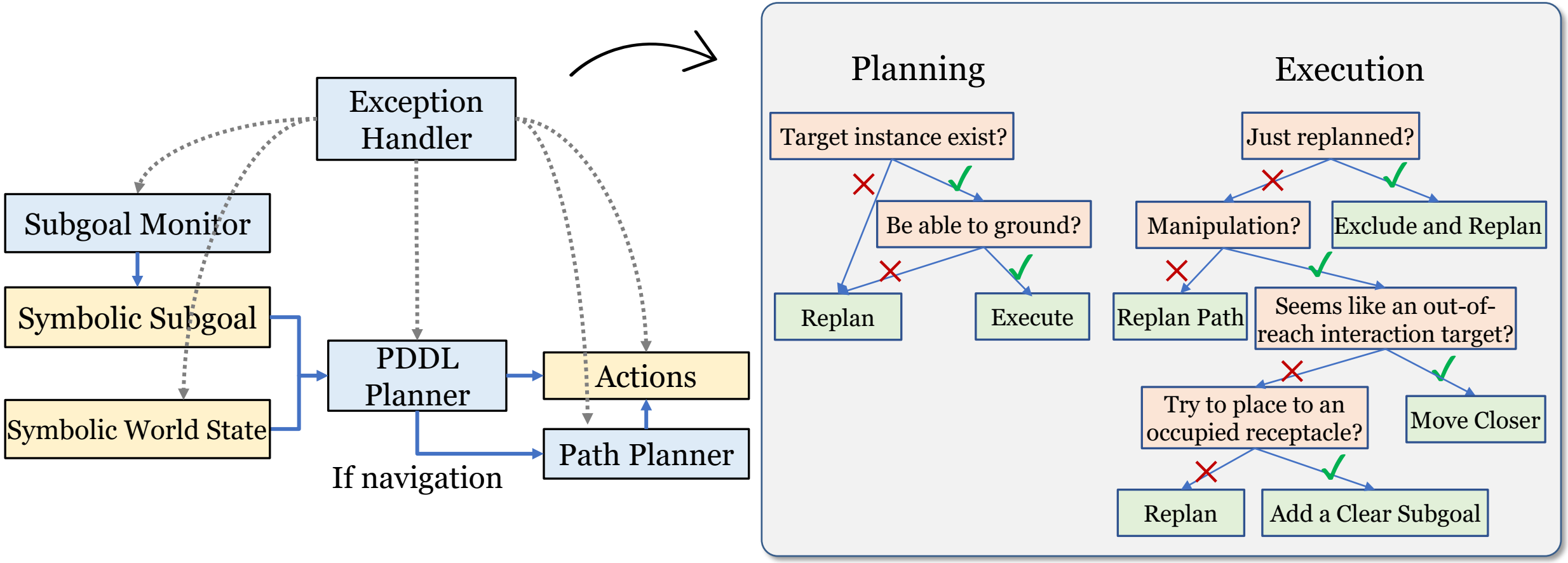

```
(:goal
    (and
        (exists (?o1 - BreadSliced ?o2 - Toaster)
            (and (parentReceptacles ?o1 ?o2))
        )
    )
)
```

Snapshot of an example subgoal "putting a piece of bread into toaster"

```
(:action Slice
    :parameters (?x - Sliceable ?y - Knives)
    :precondition (and (holding ?y) (isInteractable ?x) (not (isSliced ?x)) (not (isPickedUp ?x)))
    :effect (and
        (isSliced ?x)
        (not (isObserved ?x))
        (increase (total-cost) 15)
    )
)
```

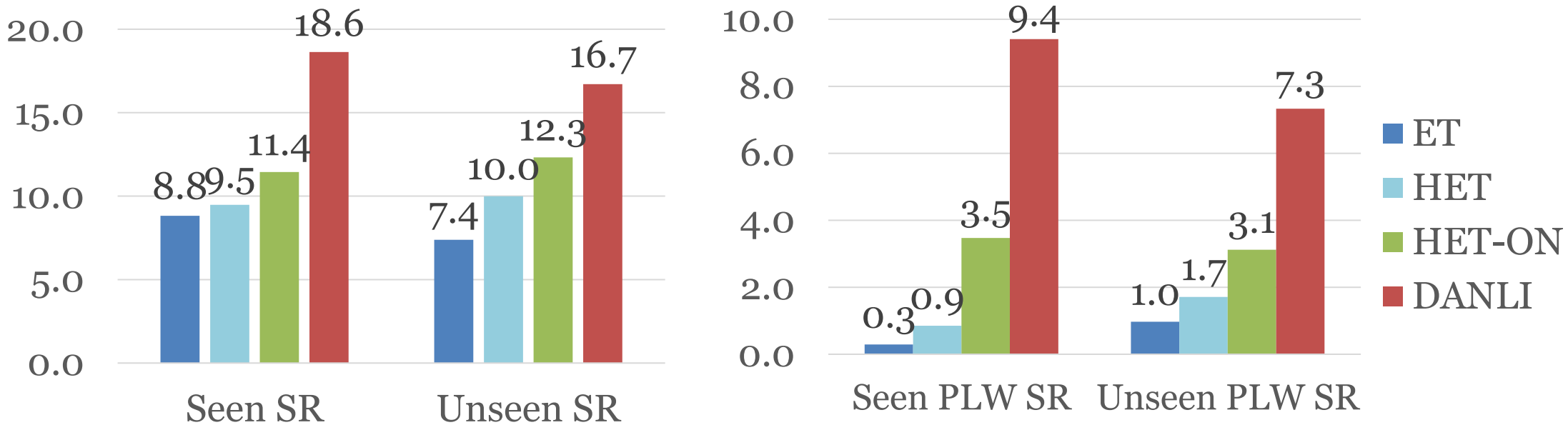
Snapshot of the action "Slice" defined in the PDDL domain

Exception Handling In Online Planning



Higher Success Rates

- DANLI outperforms reactive agents (ET, HET, HET-ON) by a large margin, especially when the task completion efficiency is considered

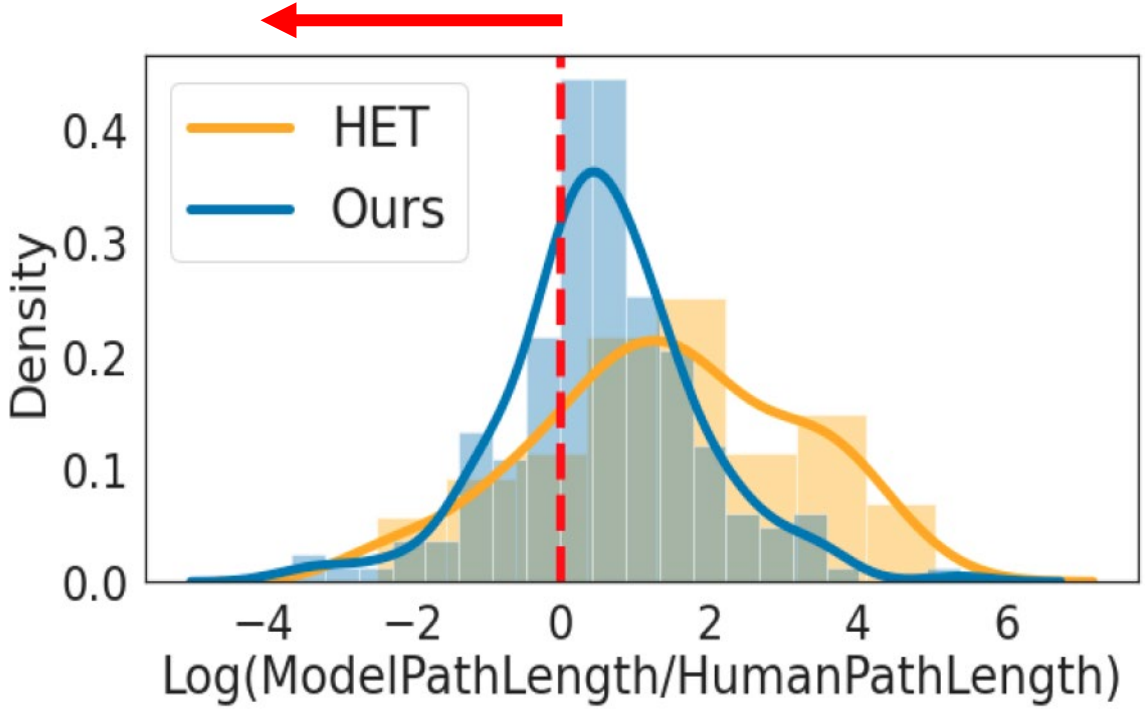


- Success Rate (SR): the proportion of successfully completed tasks
- Seen/Unseen: whether the evaluated scene is seen or unseen during training
- PLW SR: success rate weighted by the relative trajectory path length compared to humans

Higher Efficiency

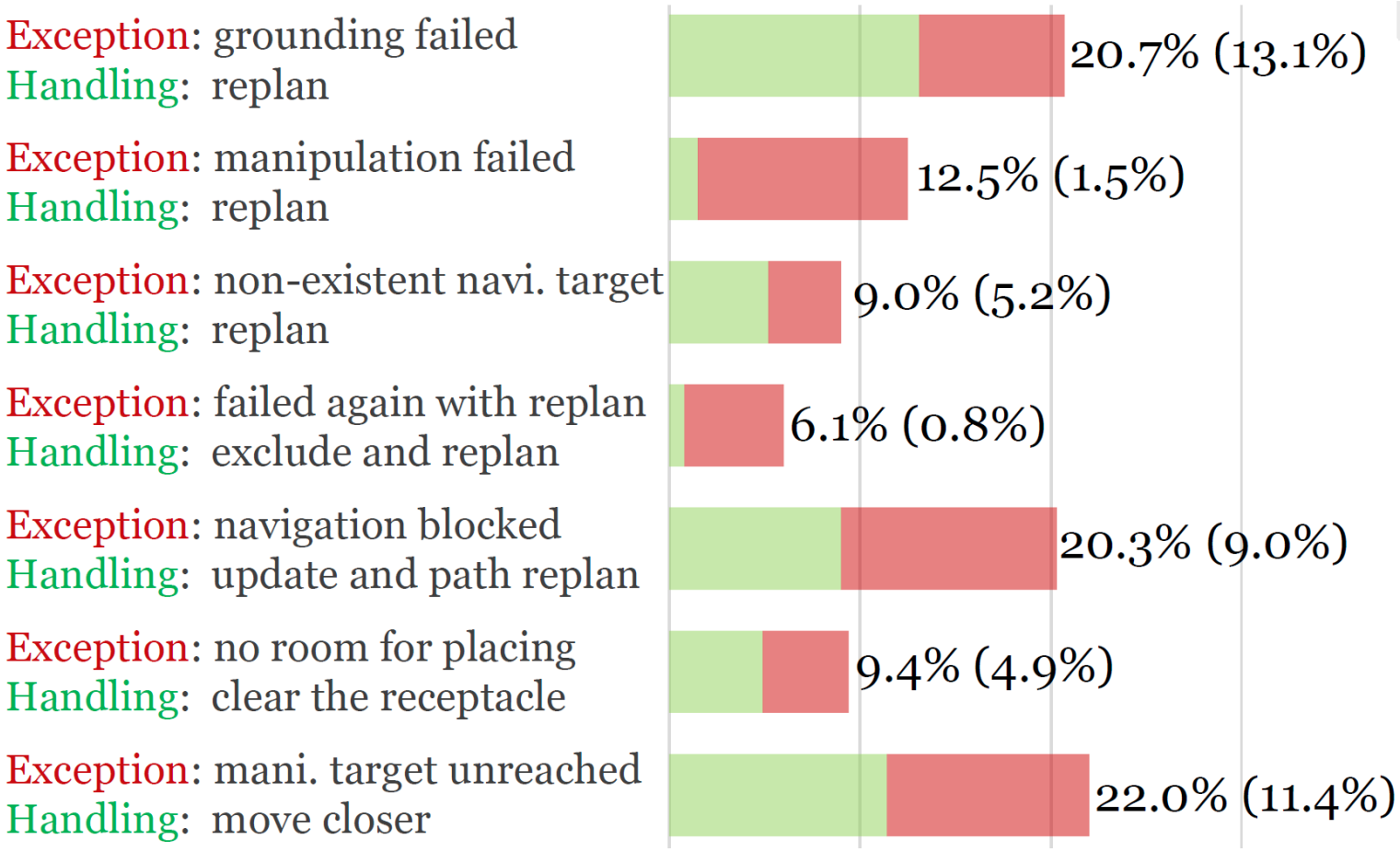
- DANLI completes tasks in fewer steps than reactive agents

Shorter path than human trajectories

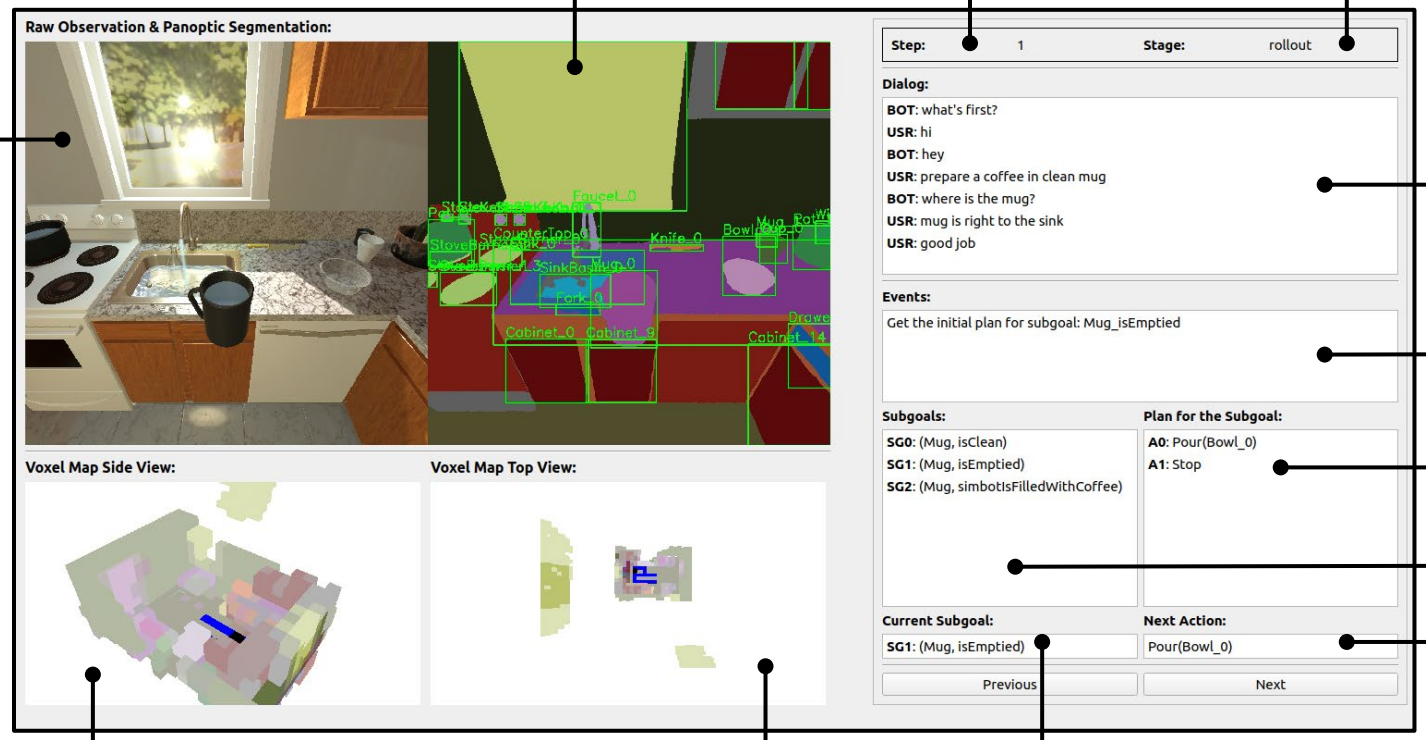


Interpretable Exception Handling

- Deliberative agent produces interpretable plans which allow fine-grain exception handling.



Demo



Panoptic Segmentation Output

Step Number

Stage: Replay/Rollout

Raw Observation

Dialog History

Key Reasoning Events

Action Plan for the Current SG

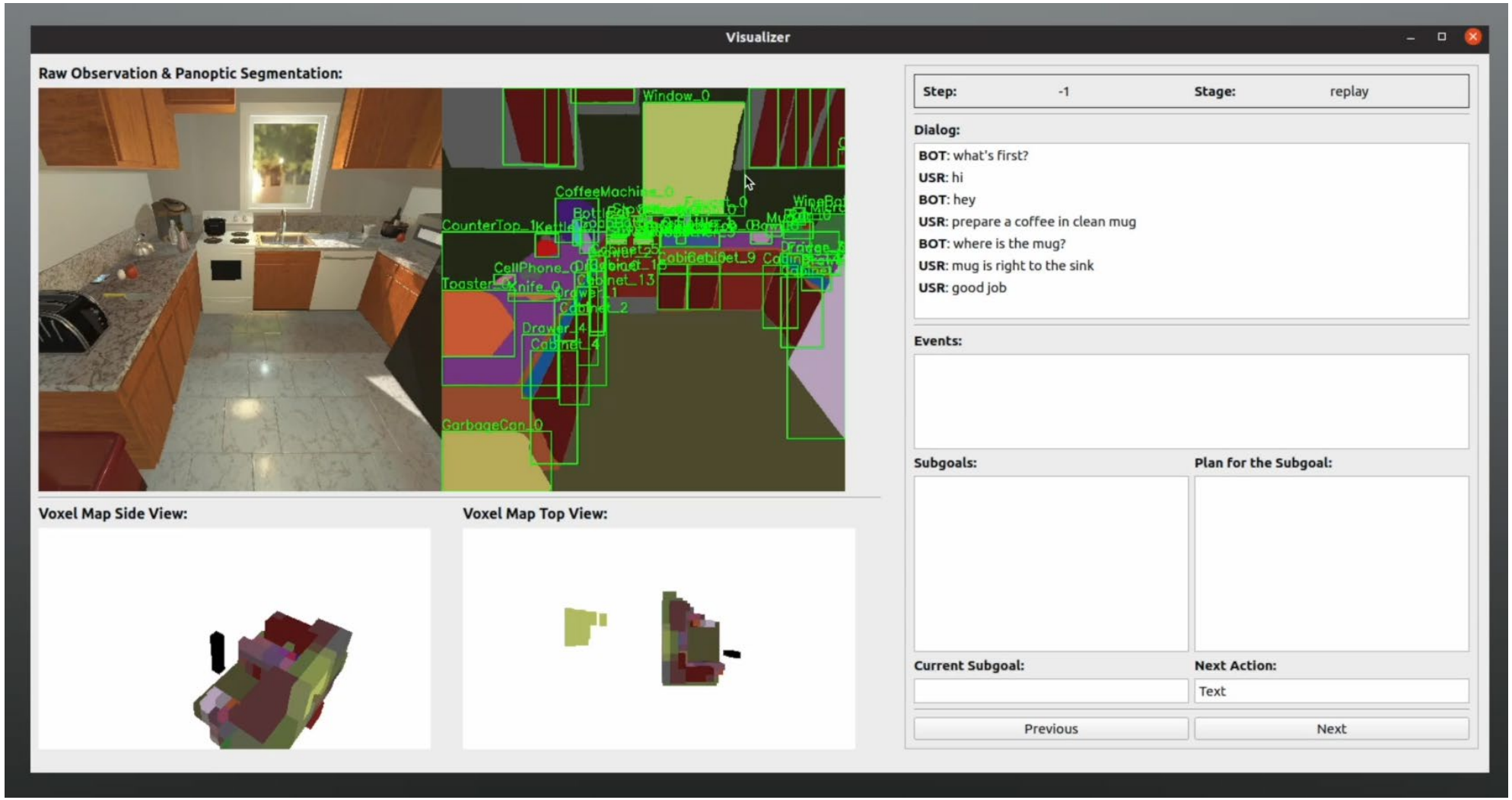
Subgoals predicted by DANLI

Voxel Map constructed by DANLI
Left: Side View / Right: Top-Down View

Current Subgoal

Next Action to Execute

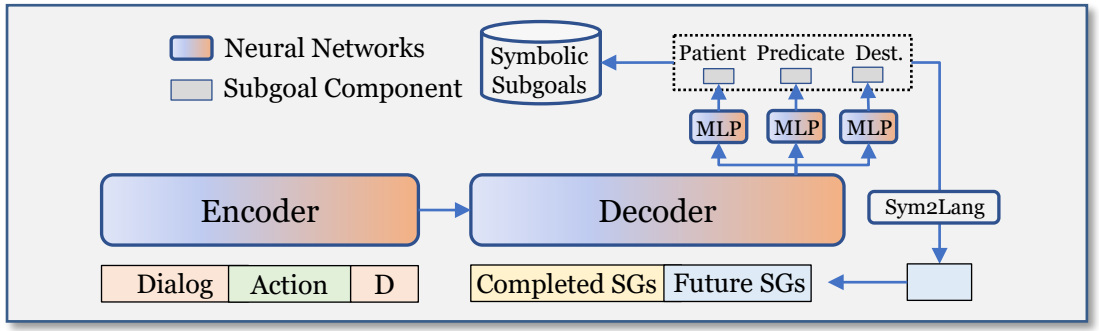
Demo


Visualizer
Raw Observation & Panoptic Segmentation:
Window_0
CoffeeMachine_0
CounterTop_1
CellPhone_0
Toaster_0
Knife_0
Drawer_4
Cabinet_4
Cabinet_13
Cabinet_9
GarbageCan_0
Voxel Map Side View:
Voxel Map Top View:
Step: -1
Stage: replay
Dialog:
BOT: what's first?
USR: hi
BOT: hey
USR: prepare a coffee in clean mug
BOT: where is the mug?
USR: mug is right to the sink
USR: good job
Events:
Subgoals:
Plan for the Subgoal:
Current Subgoal:
Next Action:
Text
Previous
Next

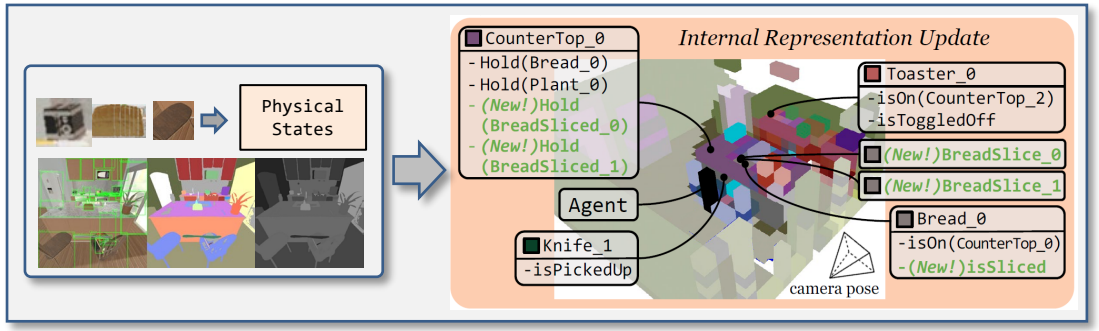
Summary - DANLI

Situated Language and Embodied Dialog (SLED) Lab

Neural Subgoal Predictor



Spatial-Symbolic World Representation



Symbolic Planner with Online Exception Handling