Hierarchical Task Learning from Language Instructions with Unified Transformers and Self-Monitoring

Yichi Zhang Joyce Y. Chai

Situated Language and Embodied Dialogue (SLED) Lab

University of Michigan, Ann Arbor, USA

Code Paper

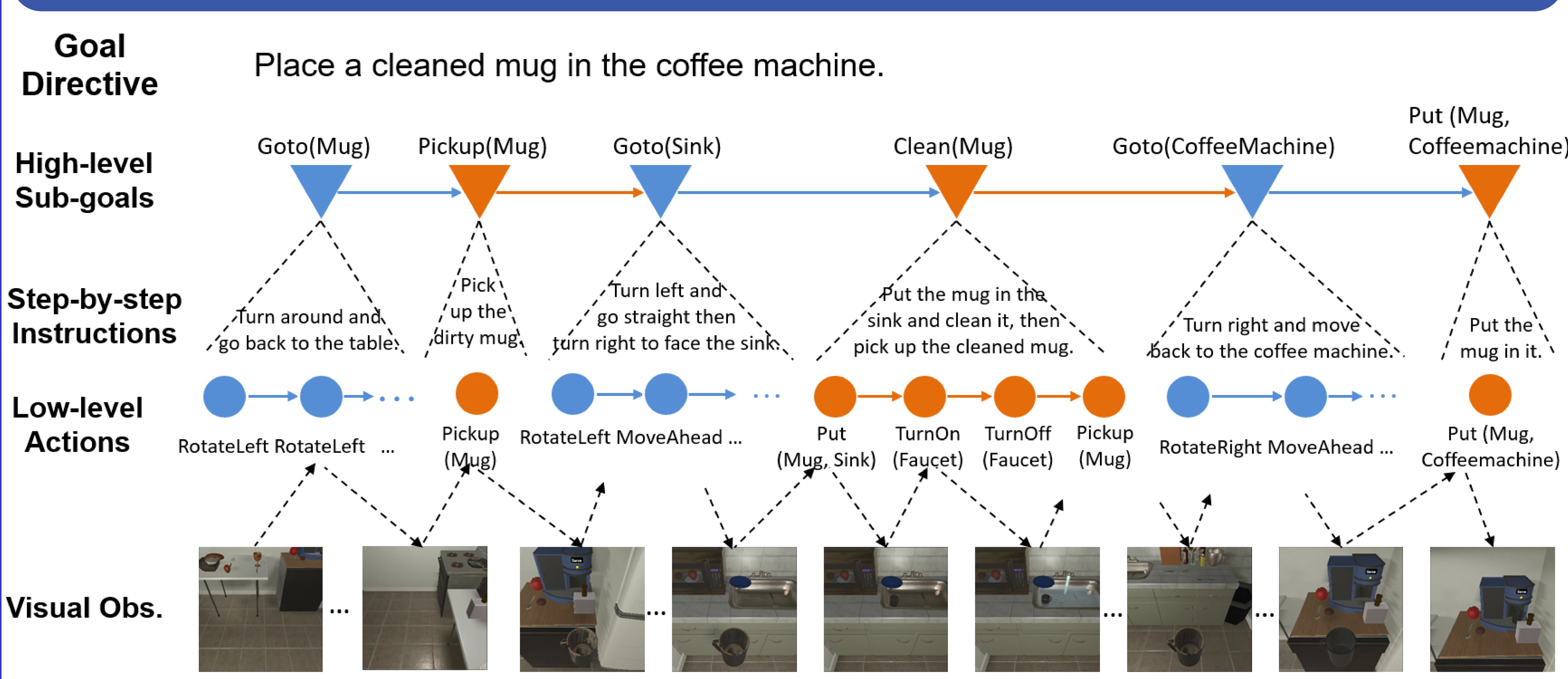
Contact: zhangyic@umich.edu

Introduction

Despite recent progress, learning new tasks through language instructions remains an extremely challenging problem. On the ALFRED benchmark for task learning, the published state-of-the-art system only achieves a task success rate of less than 10% in an unseen environment, compared to the human performance of over 90%. This paper takes a closer look at task learning for the ALFRED benchmark. The contributions include:

- Propose to decompose task learning into three sub-problems: sub-goal planning, scene navigation and object manipulation, and developed a model HiTUT (**Hi**erarchical **T**asks via **U**nified **T**ransformers) that addresses each sub-problem in a unified manner to learn a hierarchical task structure.
- HiTUT achieves new state-of-the-art result on the ALFRED benchmark (over 160% improvement on the task success rate in unseen scenes). We show that the improvement mainly sources from HiTUT's self-monitoring and backtracking ability enabled by its hierarchical task structure.
- Based on the de-composable platform, we give a more in-depth evaluation on the benchmark to better understand the complexity of its components.

An Example Task in ALFRED



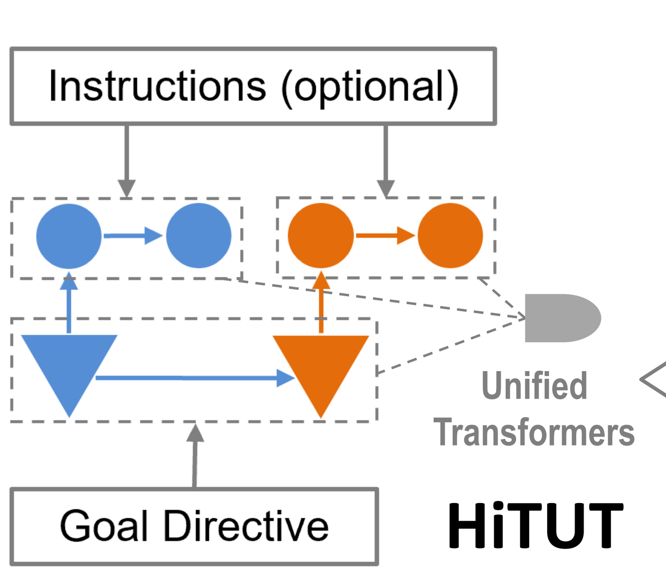
Hierarchical Tasks via Unified Transformers

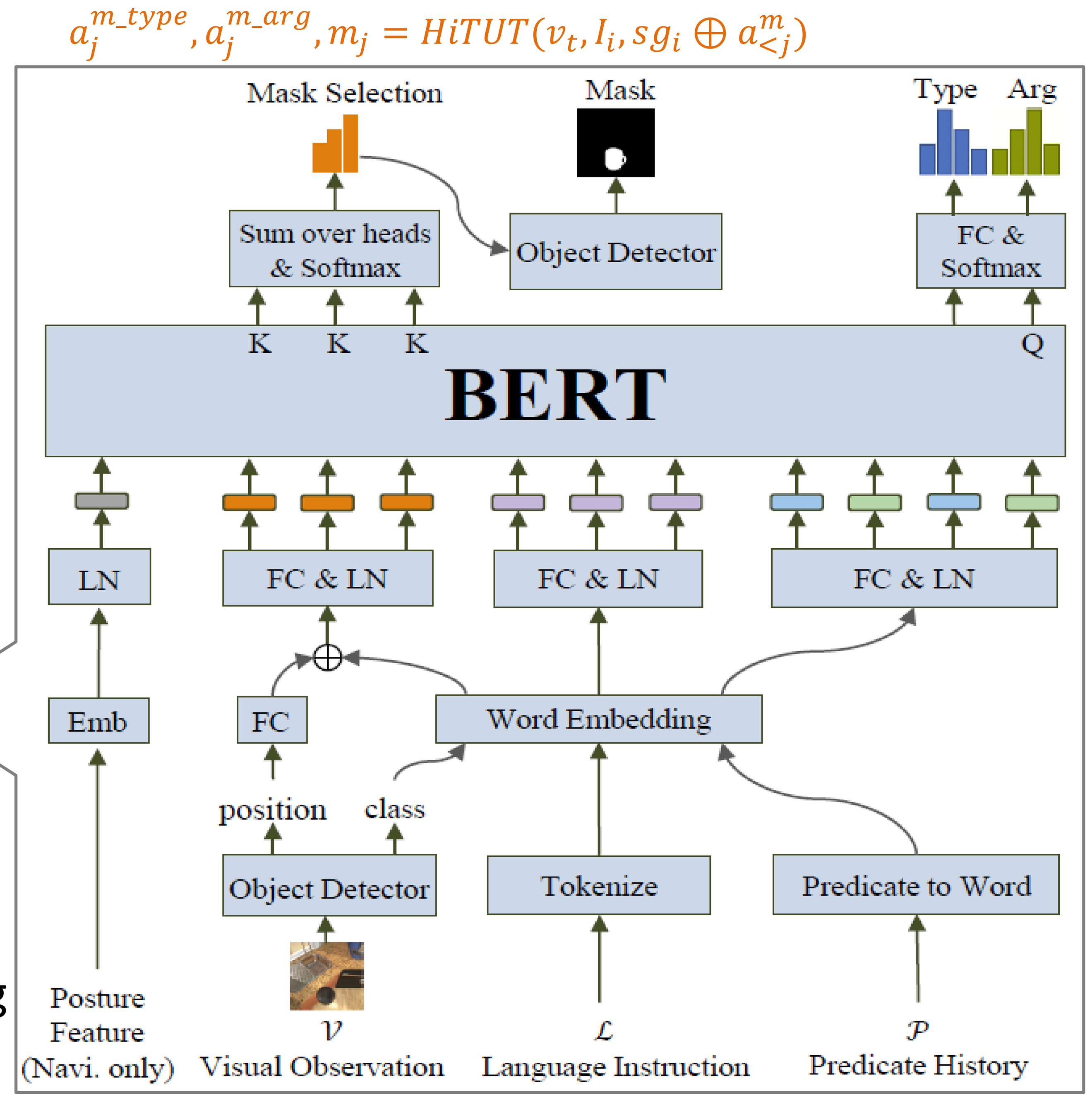
Sub-Goal Planning $sg_i^{type}, sg_i^{arg} = HiTUT(v_t, G, sg_{<i})$

Scene Navigation $a_j^{n\_type} = HiTUT(v_t, I_i, sg_i \oplus a^n_{<j})$

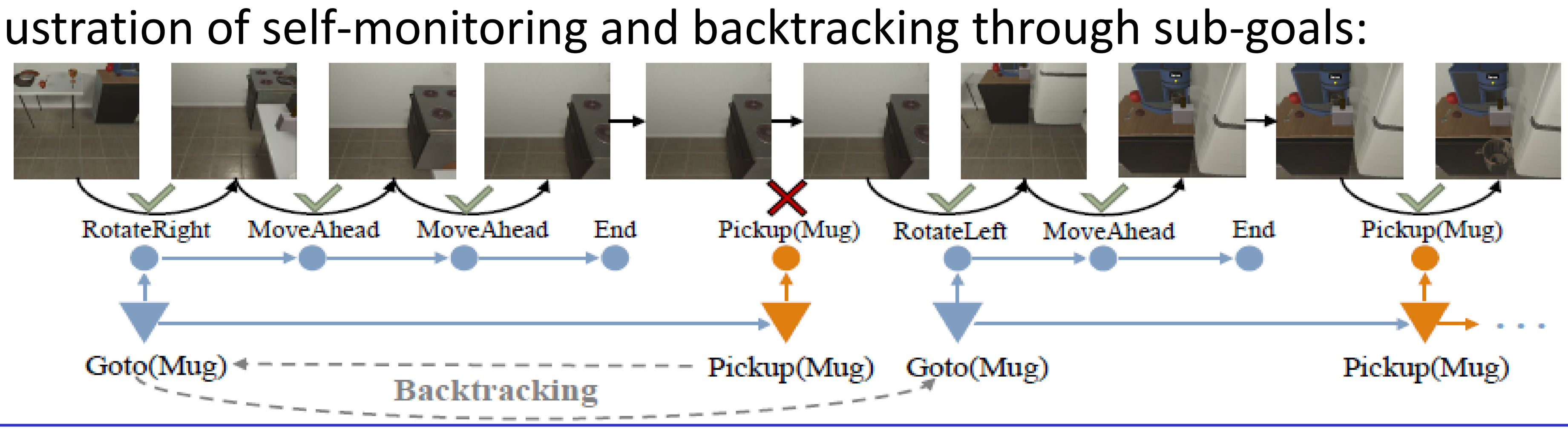
Object Manipulation $a_j^{m\_type}, a_j^{m\_arg}, m_j = HiTUT(v_t, I_i, sg_i \oplus a^m_{<j})$

- Formulation of the three sub-problems (upper)
- How the sub-problems tied up hierarchically (below)





- Trained under a multi-task manner via imitation learning from expert trajectories.
- Illustration of self-monitoring and backtracking through sub-goals:

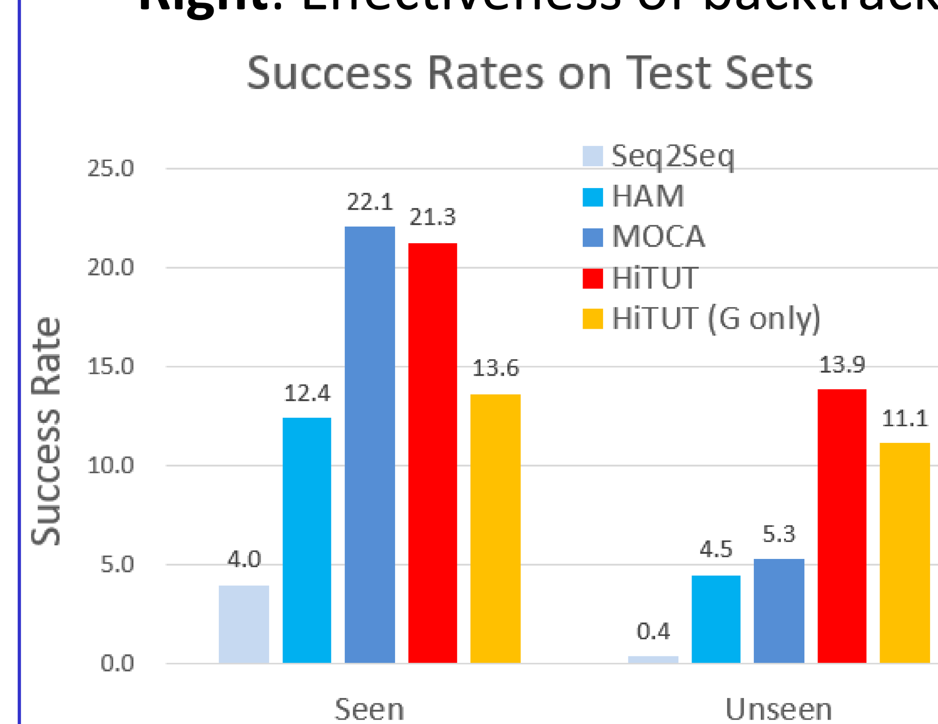


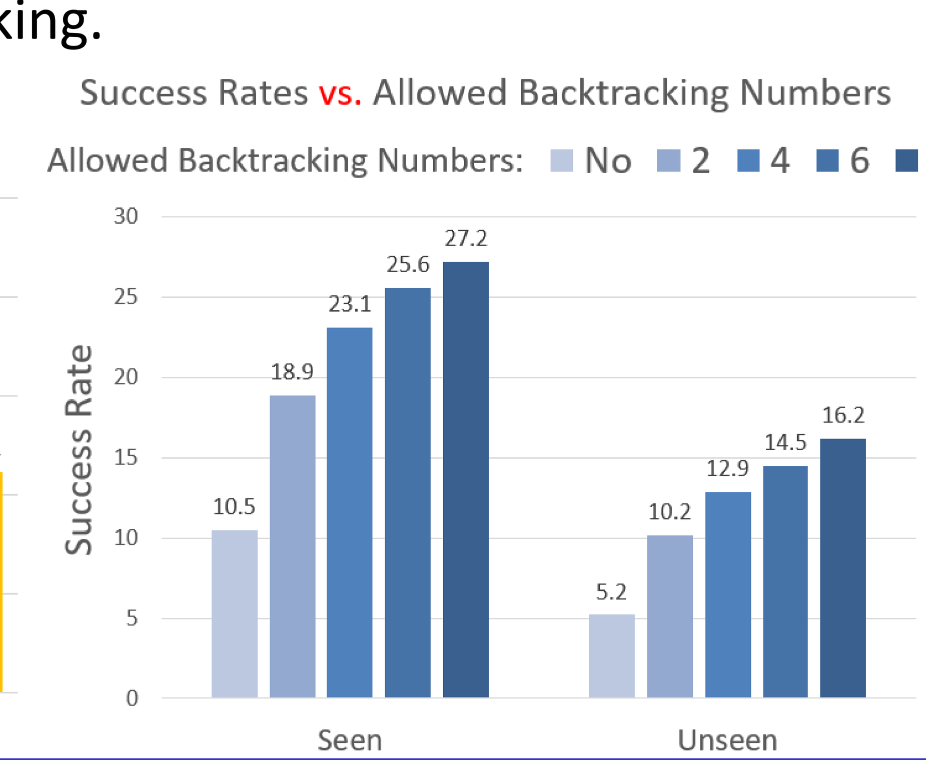
Experimental Results

- Task success rate is computed through interactive evaluation in the AI2-THOR environment. A task is considered successful if all the goal conditions (e.g. the status of mug becomes *cleaned*) are met.

Benchmark Performance

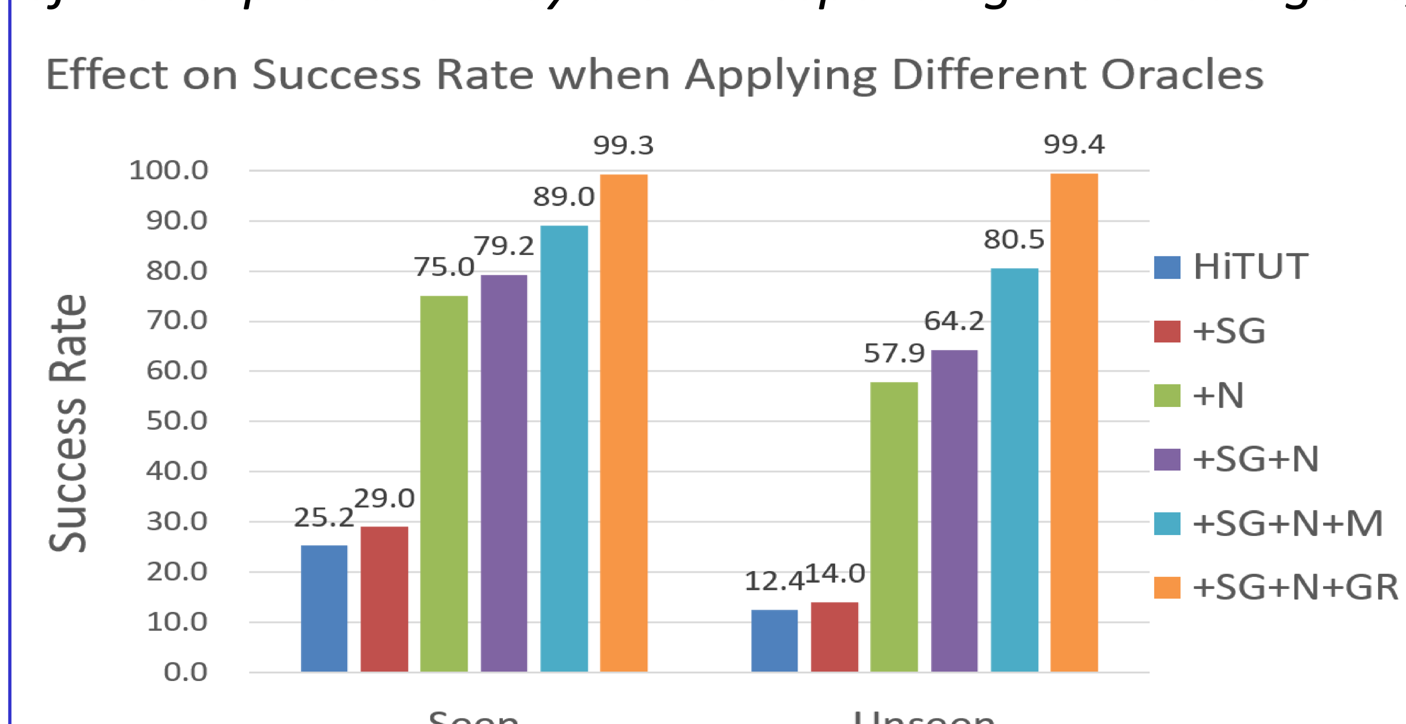
- **Left**: Overall task performance of HiTUT. In unseen scenes, HiTUT improves task success rate of 160%. Notably, HiTUT outperforms previous SOTA model (MOCA) even without step-by-step instructions.
- **Right**: Effectiveness of backtracking.





Task Complexity Analysis

Investigate how the end performance changes when replacing different part of model predictions by the corresponding oracle sub-goals/actions/masks.



(+ oracle)

SG: sub-goals

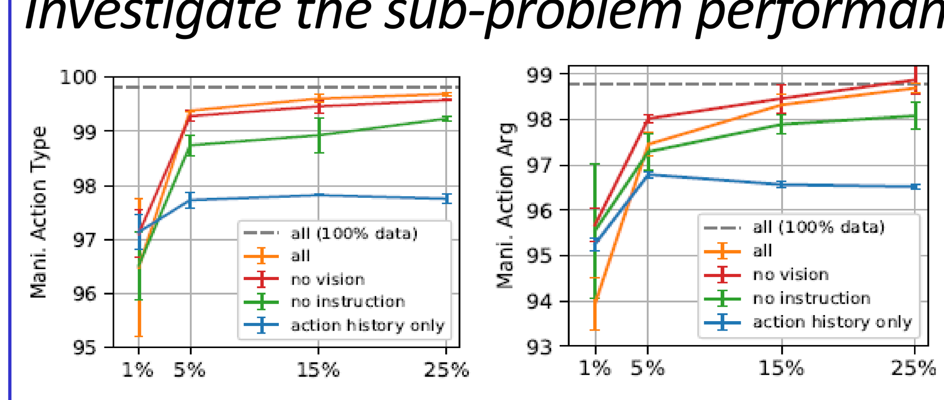
N: navigation

M: manipulation

GR: masks

- Scene navigation is the major performance bottleneck in ALFRED.
- Interactive mask generation/selection is the 2nd major cause of failure.
- Sub-goal planning and object manipulation are relatively simple.

Investigate the sub-problem performance under different resource conditions.



- Manipulation action prediction is very simple: over 96% accuracy with only 5% data conditioned on only the prediction history.
- Highly correlated manipulation actions results in shortcut of learning.